[redacted]

The New Encyclopædia Britannica

Volume 19

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
Jacob E. Safra, Chairman of the Board
Jorge Aguilar-Cauz, President

Chicago
London/New Delhi/Paris/Seoul
Sydney/Taipei/Tokyo

First Edition	1768–1771
Second Edition	1777–1784
Third Edition	1788–1797
Supplement	1801
Fourth Edition	1801–1809
Fifth Edition	1815
Sixth Edition	1820–1823
Supplement	1815–1824
Seventh Edition	1830–1842
Eighth Edition	1852–1860
Ninth Edition	1875–1889
Tenth Edition	1902–1903

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

Fourteenth Edition
© 1929, 1930, 1932, 1933, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973
By Encyclopædia Britannica, Inc.

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1997, 1998, 2002, 2003, 2005, 2007
By Encyclopædia Britannica, Inc.

4/6/2007

Printed in U.S.A.

Library of Congress Control Number: 2006921233
International Standard Book Number-10: 1-59339-292-3
International Standard Book Number-13: 978-1-59339-292-5

Britannica may be accessed at http://www.britannica.com on the Internet.

CONTENTS

Excretion and Excretory Systems

Every organism, from the smallest protist to the largest mammal, must rid itself of the potentially harmful by-products of its own vital activities. This process in living things is called elimination, which may be considered to encompass all of the various mechanisms and processes by which life forms dispose of or throw off waste products, toxic substances, and dead portions of the organism. The nature of the process and of the specialized structures developed for waste disposal vary greatly with the size and complexity of the organism.

Four terms are commonly associated with waste-disposal processes and are often used interchangeably, though not always correctly: excretion, secretion, egestion, and elimination.

Excretion. Excretion is a general term referring to the separation and throwing off of waste materials or toxic substances from the cells and tissues of a plant or animal.

Secretion. The separation, elaboration, and elimination of certain products arising from cellular functions in multicellular organisms is called secretion. Though these substances may be a waste product of the cell producing them, they are frequently useful to other cells of the organism. Examples of secretions are the digestive enzymes produced by intestinal and pancreatic tissue cells of vertebrate animals, the hormones synthesized by specialized glandular cells of plants and animals, and sweat secreted by glandular cells in the skins of some mammals. Secretion implies that the chemical compounds being secreted were synthesized by specialized cells and that they are of functional value to the organism. The disposal of common waste products should not, therefore, be considered to be of a secretory nature.

Egestion. Egestion is the act of excreting unusable or undigested material from a cell, as in the case of single-celled organisms, or from the digestive tract of multi-cellular animals. The elimination of digestive wastes is treated in the article DIGESTION AND DIGESTIVE SYSTEMS.

Elimination. As defined above, elimination broadly defines the mechanisms of waste disposal by living systems at all levels of complexity. The term may be used interchangeably with excretion.

This article discusses the eliminatory processes and mechanisms of various organisms, from the egestion of the simplest single-celled organism to the highly developed excretory process of vertebrates and of human beings. Human excretion, along with the diseases and disorders that disrupt healthy function, is explained in detail.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 421 and 423.

The article is divided into the following sections:

Elimination

BIOLOGICAL SIGNIFICANCE OF ELIMINATION

Waste disposal by unicellular and multicellular organisms is vital to their health and to the continuance of life. Animals must take in (ingest) energy-containing chemical compounds, extract a portion of the energy to power their life processes, and dispose of the unusable material or by-products formed during the energy-extraction process. An analogous series of events occurs in an internal-combustion engine. Fuel, containing energy, is taken into the engine, where it is burned, and a portion of the energy released is used to move the pistons. As in living cells, a portion of the energy-containing material (fuel) not utilized in the engine is exhausted in the form of carbon monoxide, carbon dioxide, and other by-products of combustion. Blockage of the exhaust system in an engine results in loss of efficiency and eventual total breakdown. Similarly, the rate of waste disposal in biological systems can and does provide a means of controlling the metabolic rate. Complete blockage of waste-disposal mechanisms in living systems is as effective in destroying vital functions as the cutting off of food, oxygen, or water from the system. In addition, some substances produced as metabolic by-products are toxic in themselves and must be removed from living cells at a rate equal to that at which they are produced by those cells. Thus, the excretion of waste products from living cells must occur continually in order to ensure the normal progression of vital chemical events.

Importance of elimination in living systems

Waste and poisonous substances produced by the metabolic activities of plant and animal communities must, in a similar manner, be removed or detoxified if community health is to be preserved. Collective wastes of individual organisms constituting a community, if allowed to accumulate to any marked degree, will eventually destroy the lives of all the community members.

The biosphere, composed of all individuals and communities of life forms and their environments on the Earth, is equally sensitive to the effects of waste and poison accumulation. A continual buildup of substances harmful to life forms can only result in the eventual destruction of most or all of the presently existing species of plants and animals. Humans are unique among living things in that their activities result in the production of waste materials (pollutants) that, by virtue of their chemical structure, are poisonous to all living things, including themselves. (For information about waste disposal in the biosphere, see BIOSPHERE and CONSERVATION OF NATURAL RESOURCES.)

TYPES OF WASTE: METABOLIC AND NONMETABOLIC

Waste products may be categorized as metabolic or nonmetabolic. The difference lies in whether the substances in question are produced by the chemical processes of a living cell or are merely passed through the digestive tract of an organism without actually entering into its life processes.

Nonmetabolic wastes. The nonmetabolic wastes are mainly materials that, by virtue of their chemical makeup, are indigestible or unusable by an organism. In addition, nonmetabolic wastes include any substances that are absorbed, ingested, or otherwise taken into a living system in excess of the needs and storage capabilities of the organism. These substances include digestible (metabolizable) as well as indigestible materials, and they may be excreted almost immediately, even though they are often usable as food.

Heat as a waste product

Metabolic wastes. Metabolic wastes may be separated into gases, liquids, solids, and heat. Heat, though usually not classified as a waste product, should be classified as such because it is a by-product of metabolic activity and must be eliminated to avoid harmful elevation of body temperatures in warm-blooded animals.

Gaseous wastes. Oxygen produced during photosynthetic reactions in green plants and certain bacteria may be considered to be a waste product, or at least a by-product, requiring removal. Carbon dioxide is produced by all animals and by green plants in darkness. Nitrogen gas is produced by denitrifying sulfur bacteria (*Thiobacillus*), and ammonia is excreted by decay-causing bacteria and by most invertebrate and vertebrate animals.

Liquid wastes. The sole liquid waste produced as a metabolic by-product by all animals and photosynthetic plants in darkness is water.

Solid wastes. Several important kinds of materials may be classified as solid wastes. Among them are nitrogenous wastes, by-products of protein and amino-acid metabolism by animals; nitrite and nitrate compounds produced by nitrifying bacteria; and sulfur and sulfates resulting from the metabolic activities of sulfur bacteria. Many other substances also enter into solid wastes to be disposed of by organisms. Iron compounds in an insoluble form are secreted by iron bacteria that have used soluble iron compounds. Various resins, fats, waxes, and complex organic chemicals are exuded from certain plants—as in the latex from rubber trees and milkweeds. Organic pigments from the breakdown of biological pigments, such as hemoglobin in vertebrates, become components of solid waste. Inorganic salts, including molecules and ions such as carbonates, bicarbonates, and phosphates resulting from life-sustaining chemical reactions, eventually may become solid waste products.

METHODS OF WASTE DISPOSAL

Disposal of metabolic and nonmetabolic wastes involves both active and passive mechanisms. In general, gaseous wastes are eliminated through passive mechanisms without the direct expenditure of energy on the part of the living system. The solid and liquid waste-disposal mechanisms used by higher animals are active (energy consuming) systems that separate waste materials from vital substances prior to excretion. Methods of disposal may be classified into specific and nonspecific systems.

Specific elimination mechanisms. Three pathways exist in this context: (1) the alimentary canal, (2) the respiratory system, and (3) the kidneys.

Alimentary canal. The alimentary canal is a pathway used almost exclusively for the elimination of solid wastes of an indigestible nature, and the act of elimination by this means is termed egestion. Materials disposed of in this manner have not entered the tissues of the animal but rather are the residues of enzymatic and absorptive activities occurring in the digestive tract. True metabolic wastes are excreted by means of the flow of bile from the liver into the intestine. The destruction of cells in animals produces bile pigments—residues of hemoglobin and other pigments—which may be considered to be the principal metabolic wastes eliminated via the alimentary canal. Waste disposal in this manner requires little energy expenditure other than that employed in the peristaltic contractions of muscle in the walls of the tract that act to push material along the length of the tube (see DIGESTION AND DIGESTIVE SYSTEMS).

Respiratory system. The respiratory pathway is concerned principally with the gaseous waste products of metabolism (carbon dioxide and ammonia), which move to the external environment by diffusing from the cells of origin. In invertebrate and vertebrate members of the animal kingdom, transport is by means of the circulatory system when present or simply by diffusion through the cell membranes of lower animals. A few multicellular aquatic animals lose carbon dioxide to the surrounding water by way of diffusion through the thin vascular membranes of their general body surface. In most higher animals, however, the skin is too hard or thick and nonvascular to function effectively in gas disposal. In these animals, gills and lungs—aggregations of thin, moist, vascular membranes—have evolved. Membranes of the gills of aquatic animals and the lungs of terrestrial forms are provided with large surface areas for the diffusion of waste gases from the circulatory system to the outside environment. Because carbon dioxide is soluble in the body water, it can easily diffuse into the circulatory system, in solution, from the cells of origin. Transport and excretion of carbon dioxide requires little energy as it diffuses along concentration gradients from cells to the circulation and finally to the outside environment.

Carbon dioxide elimination

Because more carbon dioxide (CO_2) is produced by metabolic activity than can be carried in the circulatory system in the form of dissolved carbon dioxide, the major portion of carbon dioxide is transported to the gills and lungs as bicarbonate (HCO_3^-), via two chemical reactions:

$$CO_2 + H_2O \leftrightarrows H_2CO_3 \rightleftarrows H^+ + HCO_3^-.$$

Thus, carbon dioxide reacts with water, producing carbonic acid (H_2CO_3), which in turn dissociates to produce a hydrogen ion (H^+) and a bicarbonate ion (HCO_3^-). In the lungs or gills, these reactions occur in the opposite direction, and carbon dioxide diffuses from the body into the outside environment. Certain aquatic animals are capable of eliminating gaseous ammonia—derived from protein breakdown—by way of specialized cells in their gill tissues.

Salt secretion via specialized gill cells occurs in marine vertebrates that constantly absorb salt through thin membranes of their oral, respiratory, and body surfaces (see RESPIRATION AND RESPIRATORY SYSTEMS).

The kidneys. Kidneys have evolved in multicellular animals as a highly sophisticated channel for waste disposal, and they function to regulate the levels of water, salts, and organic materials in the bodies of higher animals. Materials eliminated via the kidney include nitrogenous waste products (ammonia, uric acid, urea, creatine, creatinine, and amino acids), excess quantities of salts and water that may be taken into the body, and various other organic materials produced by life-sustaining chemical reactions. Functionally, the kidney is a microfilter that initially removes dissolved as well as some suspended materials from the circulatory system, along with large quantities of water. These substances are differentially reabsorbed into the blood by various kidney structures during urine formation to a degree that varies considerably throughout the animal kingdom. For example, animals that absorb large quantities of water into their bodies (such as freshwater fishes) excrete copious quantities of water in their urine. The reverse is true of many desert animals, who must conserve water and therefore produce a thick, semisolid urine. The kidney, in its various stages of evolution, functions at the expense of considerable metabolic energy and cannot be considered to be a passive system. (For a specific account of kidney structure and function, see below *Human excretion.*)

Nonspecific mechanisms of waste disposal. A multitude of disposal mechanisms exist throughout the plant and animal kingdoms for the elimination of excess plant and animal material. Among plants, the shedding and dropping of bark, leaves, and twigs might, in a broad sense, be said to represent disposal mechanisms. Certain plants, in addition, secrete or exude resins, sap, and other substances that accumulate in excessive quantities within the plant.

Elimination of particulate wastes

Specialized, mobile, amoeba-like cells exist in the blood and tissues of animals and engulf particulate wastes resulting from the disintegration of dead cells or the intake of foreign particles into the bodies of animals. Waste matter thus stored inside these small cells is removed from contact with the organism or its metabolism and may be considered to be eliminated whether or not the material is ever actually eliminated from the body of the organism during its normal life cycle.

Toxic substances are produced by normal metabolic activities. Though some of these poisons are eliminated in their original chemical form, others, such as some nitrogenous compounds, are altered biochemically to less toxic compounds. In this manner, more of the original waste may be safely stored, or permitted to accumulate without harmful effects to the organism, until it can be eliminated. In addition, toxic chemicals that are inadvertently ingested or produced by bacterial action (infection) are frequently converted to nontoxic forms by enzymatic and antibody (immune) reactions. Such materials can then be eliminated safely with other wastes along normal pathways of excretion.

Heat is eliminated from the bodies of animals by conduction to the external surface of the organism. In animals possessing a circulatory system, heat travels in its fluid from the deeper portions of the body to the surface. At the body surface, heat is lost by physical processes of convection, radiation, conduction, and evaporation of sweat.

COMPARATIVE OVERVIEW OF ELIMINATORY MECHANISMS FROM PROTISTS TO VERTEBRATES

Protista. No specialized elimination mechanisms are present in algae, fungi, protozoans, and slime molds, the main groups of protists. Metabolic wastes (carbon dioxide, water, oxygen, and nitrogenous compounds) diffuse through the cell membranes of these unicellular organisms into the outside environment. Particulate wastes pass from the bodies of certain protozoans to the exterior by way of small openings in the body surface—anal pores and other cell openings. Elimination in protists is carried out passively and therefore requires little or no expenditure of metabolic energy on the part of the organism.

Plants. Plants are not generally considered to possess special mechanisms of elimination. Photosynthetic activities of green plants, in the presence of light, produce oxygen, which diffuses out through openings in the leaves (stomata) or through the cell walls of roots and other plant structures. Excess water passes to the exterior via similar routes and is eliminated by processes of guttation (droplet exudation) and transpiration (evaporation of water from plant surfaces).

Green plants in darkness or plants that do not contain chlorophyll produce carbon dioxide and water as respiratory waste products. Carbon dioxide is secreted in the same manner as oxygen via diffusion through stomata and cell walls. Materials that are exuded by some plants—resins, saps, latexes, etc.—are forced from the interior of the plant by hydrostatic pressures inside the plant and by absorptive forces of plant cells. These forces are passive in nature, and exudation requires no energy expenditure on the part of the plant.

Animals. Diverse mechanisms have evolved that enable the various animal species to inhabit a wide range of environments. In animals whose bodies consist of a single layer of cells, waste disposal is accomplished principally by diffusion from the site of waste production to the outside environment. This method is efficient when the distances over which wastes diffuse are relatively short, when there is a high surface area to volume relationship, and when the rate of waste production is relatively low. In more complex animals, however, waste elimination by diffusion through the body wall to the exterior is less efficient because individual cells are farther removed from the exterior surface of the organism. The presence of specialized mechanisms of elimination in higher animals enables wastes to be rapidly transported to the exterior surface of the body (see below *Vertebrate excretory systems*).

Efficiency of the diffusion mechanism of waste disposal

Sponges. Phylogenetically, the sponges (phylum Porifera) are the simplest of animals. They are multicellular and composed of specialized cells, arranged in a single layer, for the maintenance of life processes. Elimination in these aquatic animals proceeds by diffusion of gaseous wastes into the surrounding water and by the ejection of solid wastes and indigestible material from the digestive cells into the streams of water that constantly flow through the animal.

Cnidarians. The jellyfishes, coral animals, ctenophores, and comb jellies have a rudimentary canallike cavity in their two-layered bodies for the ingestion, digestion, and egestion of food and wastes. Gaseous wastes are eliminated by diffusion, and solid wastes in dissolved or undissolved form pass out through an opening in the body wall that serves the dual purposes of food intake and waste elimination.

Flatworms. Flatworm bodies consist of three layers of cells, and in this aquatic group elimination is similar to that of the less complex animals. Food and solid wastes enter and leave through a common opening in the well-developed digestive tract, which consists of a mouth, pharynx, and gastrovascular cavity.

Nemertine worms. The digestive and excretory system of the aquatic proboscis worms is more efficient than that of lower animals in that a well-defined mouth, intestine, and excretory opening (anus) permit the one-way flow of food and waste through the animal. Egested food and nitrogenous wastes, which are secreted into the intestine, are passed along it to the anus by peristaltic waves of the smooth muscle lining the intestinal walls. The efficiency of waste elimination is increased by the presence of a well-defined circulatory system, which enhances the carriage of wastes to the intestine.

Nematodes. An additional excretory structure has evolved in the roundworms. Excretory canals located on both sides of the intestine facilitate waste disposal by carriage of material to an excretory pore in the body wall.

Other invertebrates. In invertebrates, increasing structural complexity is accompanied by more efficient waste-disposal mechanisms. In the phylum Mollusca (clams, snails, oysters, mollusks, octopuses, and squids), gills add another more efficient channel for waste disposal. A heart increases the rate of flow in the circulatory system and speeds the transport of wastes to the gills. An excretory, kidneylike organ removes metabolic wastes from the circulation and body fluid prior to excretion. All basic mechanisms of excretion are thus present in relatively simple animals. As invertebrates become more specialized and complex, as in the arthropods (insects, crabs, and other joint-legged animals) and annelids (segmented worms), adaptations in excretion methods allow survival in non-aquatic environments.

Vertebrates. Though the wastes produced by vertebrates differ little qualitatively from those of higher invertebrates, increased structural complexity and body size, in combination with environmental adaptations, require more specific waste-disposal mechanisms in order to maintain a constant internal environment. The presence of highly efficient, water-retaining kidneys, for example, permits vertebrates to inhabit arid, hot regions of the earth. It seems proper, within the vertebrate group, to consider elimination schemes as variations of mechanisms common to all higher animals but which enable animals to inhabit widely diversified environments. (F.C.Ke./Ed.)

General features of excretory structures and functions

The physiological process by which an organism disposes of its nitrogenous by-products is called excretion. The mechanisms for that process constitute the excretory systems, particularly such organs of vertebrate animals as elaborate and complicated as the kidney and its associated urinary ducts.

Definitions and distinctions

The meaning of excretion is most easily understood in the context of vertebrate physiology. The animal swallows food (ingestion). In the stomach and intestine some of the food is broken down into soluble products (digestion) that are absorbed into the body (assimilation). In the body these soluble products undergo further chemical change (metabolism); some are used by the body for growth, but most provide energy for the various activities of the body. Metabolism involves the uptake of oxygen and the elimination of carbon dioxide in the lungs (respiration).

Besides carbon dioxide, compounds of nitrogen arise from metabolism and are eliminated, chiefly by the kidney, in the urine (excretion). Food not digested is eliminated through the anus (defecation).

These processes are characteristic of animals in general, but not of plants. A green plant takes in carbon dioxide from the atmosphere and nitrogen (as nitrate) from the soil. It uses the energy of sunlight to build these nutrients into the materials required for growth and in the process gives out oxygen (see PHOTOSYNTHESIS).

In a broad sense animals live on plants, and the by-products of animals are the raw materials on which plants grow. These mutually supporting activities of plants and animals are kept precisely in balance by the activities of bacteria. Bacteria convert the urine and feces of animals (and also the dead bodies of both plants and animals) to carbon dioxide and nitrate. In the living world as a whole, carbon and nitrogen are in continuous circulation, driven by the energy of sunlight (see BIOSPHERE). Over most of the earth, for most of time, no by-products accumulate. Occasionally the cycles get out of balance, as they must have done during the prehistoric period when coal was being formed in the earth as a consequence of the failure of bacteria to decompose all the remains of plants.

PRODUCTS OF EXCRETION

Although every type of organism takes in some materials and eliminates others, excretion in the strict sense is a process found only in animals. For the purposes of this article excretion will be taken to mean the elimination of nitrogenous by-products and the regulation of the composition of the body fluids.

The primary excretory product arising naturally in the animal body is ammonia, derived almost entirely from the proteins of the ingested food. In the process of digestion proteins are broken down into their constituent amino acids. Some of the amino-acid pool is then used by the animal to build up its own proteins, but a great deal is used as a source of energy to drive other vital processes. The first step in the mobilization of amino acids for energy production is deamination, the splitting off of ammonia from the amino-acid molecule. The remainder is oxidized to carbon dioxide and water, with the concomitant production of the energy-rich molecules of adenosine triphosphate (ATP; see METABOLISM).

Since excessive levels of ammonia are highly toxic to most animals, they must be effectively eliminated. This is no problem in small aquatic animals because ammonia rapidly diffuses, is highly soluble in water, and escapes easily into the external medium before its concentration in the body fluids can reach a dangerous level. But in terrestrial animals, and in some of the larger aquatic animals, ammonia is converted into some less harmful compounds (detoxication). In mammals, including humans, it is detoxified to urea, which may be considered as being formed by the condensation of one molecule of carbon dioxide with two molecules of ammonia (though the biochemistry of the process is more complex than that). Urea is highly soluble in water but cannot be excreted in a highly concentrated solution because of the osmotic pressure (see below) it would exert. Because the conservation of water is important for most terrestrial animals, it is not surprising that many of them have evolved more economical methods for disposing of nitrogenous by-products. Birds, reptiles, and terrestrial insects excrete nitrogen in the form of uric acid, which is highly insoluble in water and can be removed from the body as a thick suspension or even as a dry powder.

EXCRETORY MECHANISMS

Osmotic pressure. In order to understand the advantages of the excretion of uric acid over urea it is necessary to know something about the behaviour of molecules in solution. Molecules of a solute (*e.g.*, salt, sugar) in water tend to move by diffusion from a region where they are in high concentration to one where they are in low concentration, and molecules of water tend to move in the opposite direction. If a porous membrane is placed between these regions, the movements of molecules may be variously restricted depending upon their size in relation to the size of the submicroscopic pores in the membrane. The passage of water molecules from pure water through such a membrane into a solution containing molecules that are too large to pass is called osmosis, a process that takes place spontaneously and does not require energy. This process can be reversed by applying hydrostatic pressure to the solution, a process that does require energy. The level of hydrostatic pressure at which there is no net movement of water in either direction across the membrane is called the osmotic pressure of that particular solution; the greater the concentration of dissolved molecules in the solution the greater is its osmotic pressure and the greater the force needed to remove water from it.

The osmotic process

These principles explain why more energy is required to remove water from urine containing urea than from urine containing the same weight of uric acid. The molecule of urea is smaller than that of uric acid, so that with the same weight, there are more molecules of urea to exert osmotic pressure. But an even more important difference is that whereas urea is highly soluble in water, uric acid is not. As water is progressively removed from a solution of urea, the osmotic pressure opposing further removal progressively increases. For the uric acid solution, however, as water is removed, the uric acid comes out of solution, or precipitates, when the solution is at a lower concentration, and, therefore, at a lower osmotic pressure, which does not increase further.

Regulation of water and salt balance. The mechanisms of detoxication that animals use are related to their modes of life. This is true, with greater force, of the mechanisms of homeostasis, the ability of organisms to maintain internal stability. A desert-living mammal constantly faces the problem of water conservation; but a freshwater fish faces the problem of getting rid of the water that enters its body by osmosis through the skin. At the level of the individual cell, whether it is the cell that constitutes a unicellular organism or a cell in the body of a multicellular organism, the problems of homeostasis present themselves in similar ways.

To continue its intracellular processes a cell must maintain an intracellular chemical environment in which the concentrations of various ions (see below) are kept constant in the face of changing concentrations in the medium surrounding the cell. This is the task of the cell membrane. In the higher animals the task is easier since cells in the interior of their bodies are bathed in an internal medium—the blood—whose composition is regulated so as to minimize the effects of changes in the external medium. This regulatory function is undertaken by specialized cells or organs such as the kidney, thereby lessening the regulatory burden of the other cells of the body.

The biological necessity for homeostatic mechanisms is particularly urgent for controlling the inorganic components of cells and body fluids. Inorganic salts can exert even greater osmotic pressure against membranes impermeable to them than urea. This is so because, under the conditions in the body, they are almost completely dissociated into their component ions. For example, a molecule of common salt (sodium chloride) is dissociated into two inorganic ions—a positively charged sodium ion and a negatively charged chloride ion—both of which can exert osmotic pressure.

Aside from their osmotic effects, inorganic ions have profound effects upon metabolic processes, which in general will take place only in the presence of appropriate concentrations of these ions. The most important inorganic ions in organisms are the positively charged hydrogen, sodium, potassium, calcium, and magnesium ions, and the negatively charged chloride, phosphate, and bicarbonate ions. The membranes of cells are not completely impermeable to these ions and are in fact endowed with the ability to transport ions between the inside and outside of the cell, whereby they control the concentrations of ions within the cells; when such transport is in the direction that requires a supply of energy, it is called active transport (see CELLS: *The plasma membrane*).

Active transport of ions

Osmotic regulation is the maintenance of the normal concentration of the body fluids; *i.e.*, the total concentration of all dissolved substances (solutes) that would exert osmotic pressure against a membrane impermeable to them. Osmotic regulation controls the amount of water in the body fluids relative to the amount of osmotically active solutes. Ionic regulation is the maintenance of the concentrations of the various ions in the body fluids relative to one another. There is no consistent distinction between the two processes; organs that participate in one process at the same time participate in the other.

Principal excretory structures. Whereas the kidney is the principal organ subserving both nitrogenous excretion and osmotic and ionic regulation in the mammalian body, these functions are not always performed by a single organ in other animals. As indicated earlier, primitive aquatic animals do not require any special provision for nitrogenous excretion. But by reason of their permeable skins they may have serious problems of osmotic and ionic regulation, especially in fresh water, where cells covering the surface of the body have the ability to actively transport salts into or out of the animal. In some cases these nonkidney regulatory activities are performed by certain specialized cells; *e.g.,* in the gills of fishes (see below). In other cases, specialized cells are assembled into organs of salt uptake or salt elimination; *e.g.,* the salt glands of birds (see below).

This dispersal of the regulatory function may be the primitive condition, for it is only in the more highly evolved terrestrial animals that the regulatory function is restricted to an excretory system proper. This is readily understandable in view of the need of terrestrial animals to conserve water. This evolutionary development toward one system reaches its climax in the birds, reptiles, and terrestrial insects, in which all the processes of elimination that might involve loss of water—defecation, nitrogenous excretion, and ionic regulation—converge upon the same final channel.

For the excretory organs of a wide variety of vertebrate and invertebrate animals, there is evidence that the primary process of urine production is nonselective, in that in those animals all substances dissolved in their body fluids, with the possible exception of proteins, are found in the primary urine. In many animals the primary urine is produced by filtration from the blood. At a later stage, substances in the primary urine that are useful to the body are selectively reabsorbed. In addition, a few substances are known to be actively transported (secreted) into the urine.

The nonselective formation of primary urine serves another aspect of excretion: the elimination of foreign substances. Mechanisms of active transport are highly specific to the substances transported. All dissolved constituents of the body fluids pass freely into the primary urine, and then specific reabsorptive mechanisms gather up the "wanted" substances. In this way a natural economy automatically eliminates "unwanted" substances simply by not providing mechanisms for their reabsorption.

INVERTEBRATE EXCRETORY SYSTEMS

Regulation of body fluids

In their detoxication mechanisms, so far as they have been investigated, the invertebrates in general conform to the principles applying to all animals, namely, that aquatic forms get rid of ammonia by diffusion through the surface of the body; terrestrial forms convert ammonia to uric acid. This implies that in aquatic forms the excretory organ is principally of importance for the composition of their body fluids. Normally, the body fluids of marine invertebrates have the same concentration as seawater; they usually differ, however, in the proportions of ions, with relatively more potassium and less magnesium than seawater. Furthermore, their urine normally has the same concentration as seawater, but correspondingly it contains less potassium and more magnesium. In freshwater invertebrates the urine is commonly, though not invariably, more dilute than the body fluids. By producing dilute urine a freshwater invertebrate conserves the salt content of its body while eliminating the water that enters its body by osmosis through its water-permeable surface.

Some invertebrates, notably echinoderms, cnidarians, and sponges, have no organs to which an excretory function can be confidently ascribed. Since all of these animals are aquatic, it is reasonable to suppose that they excrete nitrogen (as ammonia) by simple diffusion. Their body fluids (where present) are closely similar to seawater in composition, and it may be presumed that regulation operates only at the cellular level.

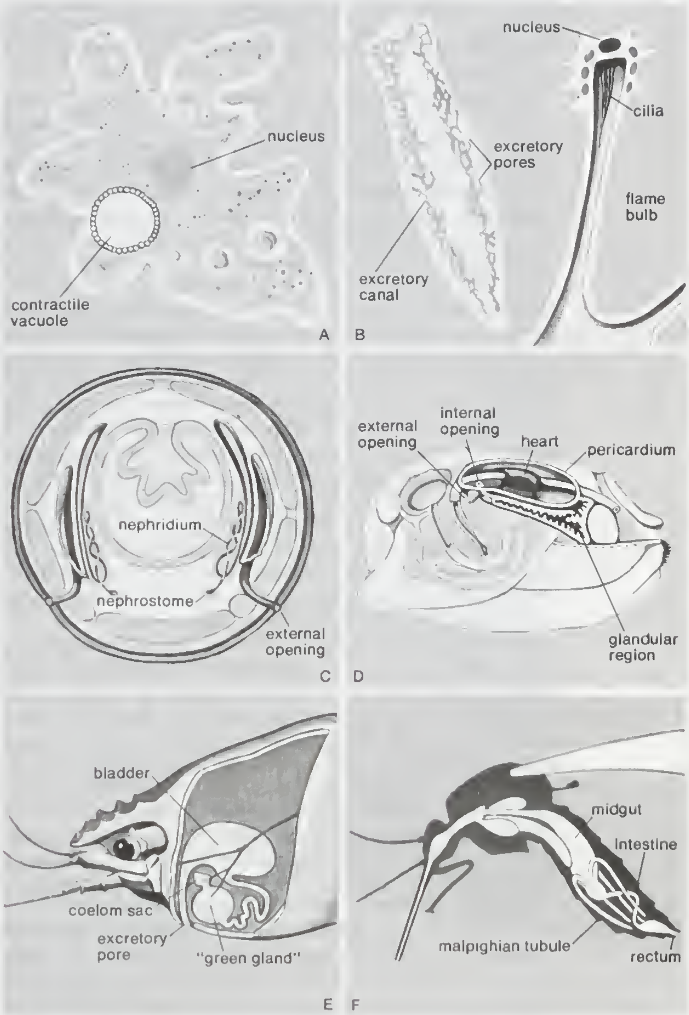


Figure 1. *Invertebrate excretory systems.*
(A) Contractile vacuole of an amoeboid protozoan. (B) Protonephridial system of a flatworm, with enlargement of a single-celled flame bulb that terminates the tubules of excretory canals. (C) Metanephridial system of an earthworm, with paired nephridia in one segment. (D) Renal organ of a clam, cut away to show glandular region. (E) Renal system of a crayfish, exposed through cut-away shell. (F) Excretory system of a mosquito.

The excretory organs of other invertebrates are of diverse evolutionary origin. This is not to say, however, that each invertebrate phylum has evolved its own particular type of excretory organ; rather, there appear to be five main types of invertebrate excretory organ: contractile vacuole, nephridium, renal gland, coxal gland, and malpighian tubule.

The contractile vacuoles of protozoans. Some protozoan animals possess an organelle having the form of an internal sac, or vacuole, which enlarges by the accumulation of a clear fluid and then discharges its contents to the exterior. The cycle of filling and emptying may be repeated as frequently as every half minute. The chief role of the contractile vacuole appears to be in osmotic regulation, not in nitrogen excretion.

Contractile vacuoles occur more frequently and are more active in freshwater species than in closely related marine species. In fresh water, the concentration of dissolved substances in the cell is greater than in the external medium, and the cell takes in water by osmosis. If the contractile vacuole is put out of action, the cell increases in volume. If the concentration of salts in the medium increases—which would have the effect of decreasing the rate of osmosis—the rate of output by the contractile vacuole diminishes. The fluid eliminated by the vacuole is more dilute than the cytoplasm.

The nephridia of annelids, nemertines, flatworms, and rotifers. The word nephridium applies in its strict sense only to the excretory organs of annelids, but it may usefully be extended to include the excretory organs of other phyla having similar characteristics. Annelids are segmented animals that typically contain a pair of nephridia on each segment. Each nephridium has the form of a very fine tubule, often of considerable length; one end usually opens into the body cavity and the other to the exterior. In some annelids, however, the tubule does not open into the body cavity but ends internally in a cluster of cells of a special type known as solenocytes, or flame cells. The possession of solenocytes by some annelids is one of the characteristics that allies them with other nonsegmented phyla that have no true body cavity. They also have a system of tubules opening at the surface and ending internally in flame cells embedded among the other cells of the body. In most cases, there is no regular arrangement of the various parts of the system. Animals belonging to all of these phyla are primarily aquatic, and, in the few cases known, the main excretory product is ammonia. How much of it leaves the body by the nephridia and how much through the body surface is not known.

Few physiological studies have been made on nephridia other than those of the earthworm. Although the earthworm is considered a terrestrial animal, its relationships with its environment are characteristically those of a freshwater animal. The nephridium of the earthworm is longer and more complex than that of marine annelids, four regions being distinguishable. Body fluid enters the nephridium via an internal opening called the nephridiostome. As the fluid passes along the tubule, probably driven by cilia, its composition is modified. In the two lower regions of the tubule the fluid becomes progressively more dilute, presumably as a result of the reabsorption of salts. Finally, a very dilute urine passes into the bladder (an enlarged portion of the tubule) and then to the exterior through the external opening, or nephridiopore. The rate of urine flow for an earthworm may be as much as 60 percent of its body weight in a period of 24 hours.

The common plan of the renal gland

The renal glands of mollusks. The anatomical form of the renal gland varies from one class of mollusks to another, but a common plan is clearly evident. The renal gland is a relatively wide tube opening from a sac (the pericardium) surrounding the heart, at one end, and to the mantle cavity (effectively to the exterior) at the other. There is a single pair of renal glands; in some forms one member of the pair may be reduced or absent. Clams have the simplest arrangement; the region nearest to the pericardium has glandular walls and gives way to a nonglandular, wider tube that extends to the urinary opening.

The vast majority of mollusks are aquatic and excrete nitrogen in the form of ammonia. In octopuses, however, nitrogen is excreted as ammonium chloride, which is quite strongly concentrated in the urine. Terrestrial snails and slugs excrete uric acid but may also excrete ammonia when living in moist surroundings.

In all mollusks so far investigated the primary process in urine production appears to be filtration of the blood. This may take place through the wall of the heart into the pericardium, or from blood vessels that supply the glandular part of the renal gland. The composition of the primary urine may be altered by reabsorption or secretion, or both. In freshwater mollusks salts are reabsorbed in the glandular tube and in the wide tubule, and the final urine is more dilute than the blood. The rate of urine flow is high, up to 45 percent of the body weight per day in the freshwater mussel. In marine mollusks the urine has the same concentration as the blood, but (in the few cases examined) its ionic composition is different.

The coxal glands of aquatic arthropods. Coxal glands are tubular organs, each opening on the basal region (coxa) of a limb. Since arthropods are segmented animals, it is reasonable to suppose that the ancestral arthropod had a pair of such glands in every segment of the body. In modern crustaceans there is, as a rule, only a single pair of glands, and in higher crustaceans these open at the bases of the antennae. Each antennal gland is a compact organ formed of a single tubule folded upon itself. When unraveled the tubule is seen to comprise three or four easily recognizable regions. The tubule arises internally as a small sac, the coelomic sac, which opens into a wider region, the labyrinth, having complex infoldings of its walls. The labyrinth opens either directly into the bladder, as in marine lobsters and crabs, or into a narrow part of the tubule, the canal, which in turn opens into the bladder, as in freshwater crayfishes.

Ion concentration in urine

The coelomic sac, well supplied with blood vessels, gives evidence that the primary process in urine production is filtration of the blood through the wall of the coelomic sac in a manner analogous to filtration in the glomerulus and Bowman's capsule of the vertebrate kidney (see below). In lobsters and marine crabs the urine in all parts of the organ has the same ion concentration as the blood. In freshwater crayfishes the urine has the same concentration as far as the end of the labyrinth; from that point on reabsorption takes place in the canal and the urine leaves the body as a very dilute solution. The addition of the canal to the system demonstrates one way crustaceans have adapted to life in fresh water. But this is not the only way in which the regulatory problem is solved in freshwater crustaceans. In freshwater crabs, for example, there is a great decrease in the water permeability of the surface (principally the gills) so that water enters by osmosis quite slowly. In contrast to the rate of urine flow in a freshwater crayfish (about 5 percent of the body weight per day), that of the freshwater crab is 100 times less (about 0.05 percent). In the crab the urine has the same concentration as the blood, but because the flow is so small the salt loss via the urine is negligible. A few semiterrestrial crabs are known to produce urine more concentrated than the blood.

In all crustaceans for which analyses are available the concentrations of ions in blood and urine differ. At a urine flow of 5 percent of the body weight per day the activities of the antennal glands are certainly capable of effecting changes in the composition of the blood. These activities are somehow coordinated with salt uptake by the cells of the body surface so as to subserve homeostasis. The role of the antennal glands in nitrogenous excretion seems to be unimportant.

The malpighian tubules of insects. Although some terrestrial arthropods (*e.g.,* land crabs, ticks) retain the coxal glands of their aquatic ancestors, others, the insects, have evolved an entirely different type of excretory system. The malpighian tubules, which vary in number from two in some species to more than 100 in others, end blindly in the body cavity (which is a blood space) and open not directly to the exterior but to the alimentary canal at the junction between midgut and hindgut. The primary urine issuing from the malpighian tubules has to pass through the rectum before it leaves the insect's body, and in the rectum its composition is markedly changed. The insect excretory system therefore comprises the malpighian tubules and the rectum acting together.

Primary urine formation in insects

The malpighian tubules are bathed in the insect's blood, but since they are not rigid it is impossible for any hydrostatic pressure to be developed across their walls, such as could bring about filtration. The primary urine is formed by a process of secretion in the following way: Potassium ions are actively transported from the blood into the cavity of the tubule and are necessarily followed by negatively charged ions so as to maintain electroneutrality. In turn, water follows the ions, probably by osmosis, and various other substances—sugars, amino acids, and urate ions—also enter the primary urine by diffusion from the blood.

The primary urine, together with soluble products of digestion and insoluble indigestible matter from the midgut, then passes to the rectum. There (or in some insects at an earlier stage) the urine is acidified and the soluble urate is thereby converted to insoluble uric acid, which comes out of solution. Water is then reabsorbed together with the soluble products of digestion and other useful substances, including the bulk of the ions that entered the primary urine. In insects that live in dry surroundings the rectum has remarkable powers of reabsorption, its contents finally being voided as hard, dry pellets containing solid uric acid.

The activity of the excretory system in insects is under hormonal control. This has been most clearly demonstrated in the case of *Rhodnius,* a bloodsucking bug. Immediately after the ingestion of a blood meal there is a rapid flow of urine whereby most of the water taken in with the blood meal is eliminated. The distension of the body after ingestion is the stimulus that causes certain cells in the central nervous system to release a hormone that acts upon the malpighian tubules to promote a brisk flow of primary urine.

VERTEBRATE EXCRETORY SYSTEMS

The kidney and its associated ducts are the excretory system of the mammal, and, as already noted, most of the nitrogenous waste arising in the mammalian body is excreted as urea. Other nitrogenous compounds regularly present in the urine in smaller amounts are uric acid (or the closely related compound allantoin) and creatinine; both of these arise mainly as by-products of the renewal and repair of tissues.

In birds, reptiles, and amphibians the kidneys are compact organs, as they are in mammals, but in fishes they are narrow bands of tissue running the length of the body (see below under *Evolution of the vertebrate excretory system*). In amphibians, as in mammals, the main excretory product is urea. In birds and reptiles it is uric acid. In most fishes the main excretory product is ammonia.

Mammals. The mammalian kidney (Figure 2) is a compact organ with two distinct regions: cortex and medulla. The functional unit of the kidney is the nephron. Each nephron (Figure 2) is a tubular structure consisting of four regions. It arises in the cortex as a small vesicle about one-fifth of a millimetre (0.008 inch) in diameter, known as Bowman's capsule, into which projects a tuft of capillary blood vessels, the glomerulus. Bowman's capsule is continuous with the proximal convoluted tubule, which also lies in the cortex. Following the proximal convoluted tubule is the loop of Henle, which descends into the medulla and then runs straight up again to the cortex where it continues as the distal convoluted tubule. A collecting tubule, into which several nephrons open, courses through the medulla to open a wide cavity, the pelvis of the kidney. From the pelvis the ureter leads to the bladder, and from the bladder the urethra leads out of the body.

From J M Forrester, R Passmore, and J S Robson (eds), *A Companion to Medical Studies*, vol 1, 3rd ed (1985), Blackwell Scientific Publications Ltd

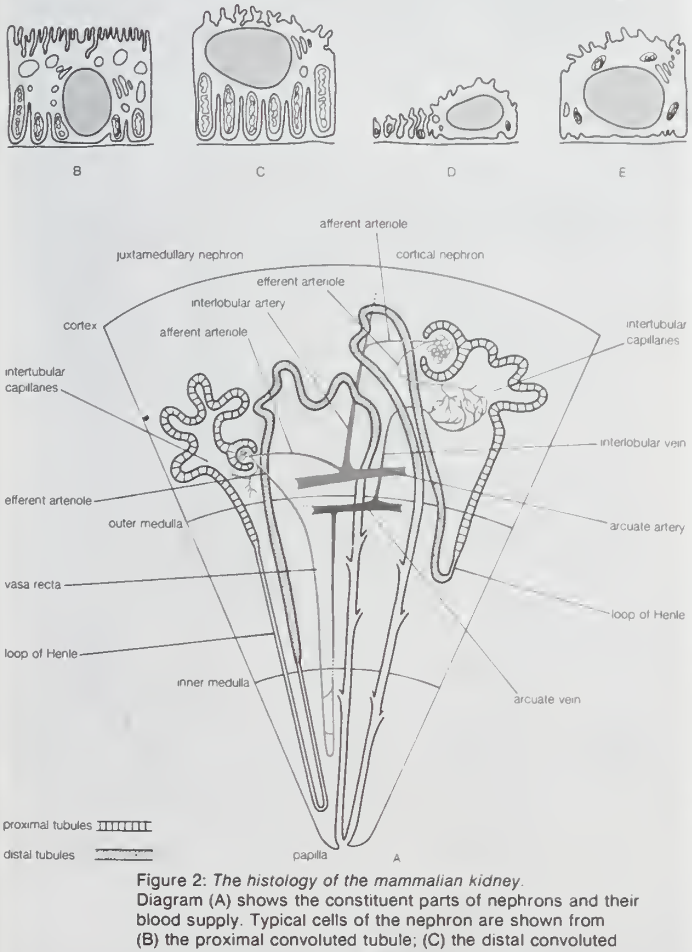


Figure 2: *The histology of the mammalian kidney.*
Diagram (A) shows the constituent parts of nephrons and their blood supply. Typical cells of the nephron are shown from (B) the proximal convoluted tubule; (C) the distal convoluted tubule; (D) the thin limb of the loop of Henle; and (E) the collecting duct. In each case the cell's luminal surface is at the top and the external surface is at the bottom.

Mechanism of urine formation in mammals

The mechanism of urine formation involves three processes: filtration, reabsorption, and secretion. Primary urine is formed by filtration from the blood. From this primary urine certain substances are reabsorbed into the blood and other substances are secreted into the primary urine from the blood. The word secretion is used by renal physiologists to imply transport, other than by filtration, from the blood to urine. Filtration implies that all molecules below a certain size are allowed to pass nonselectively into the primary urine; reabsorption and secretion imply the existence of specific mechanisms for the transport of specific substances.

The membrane covering the glomerulus allows the passage of water and all the constituents of the blood plasma except proteins. The glomerular capillaries are intercalated in the course of an artery, with the consequence that the pressure of the blood in these capillaries is higher than in the capillaries in other parts of the kidney. Opposed to the blood pressure are the pressure of the fluid within Bowman's capsule and the osmotic pressure exerted by the proteins of the blood plasma; but the blood pressure is sufficiently in excess of the sum of these to ensure a rapid flow of fluid, the glomerular filtrate or primary urine, into Bowman's capsule. The glomerular filtrate contains the nitrogenous compounds ultimately to be excreted in the urine. As the glomerular filtrate passes through the proximal tubule, 80 percent of the water, and many substances of value to the body (*e.g.,* glucose), is reabsorbed into the blood capillaries surrounding the tubule. This reabsorptive process is accomplished without any change in the concentration of the tubular fluid, which remains the same as that of the blood plasma.

After traversing the loop of Henle, the remaining 20 percent of the glomerular filtrate passes into the distal tubule, where further reabsorption, notably of salts, takes place. If this is accompanied by a proportionate reabsorption of water, the tubular fluid remains at the same concentration as the blood plasma, but if the reabsorption of water is restricted, as it may be in certain circumstances (see below), the tubular fluid becomes more dilute than the blood plasma. Under normal physiological conditions some 15 percent of the glomerular filtrate is reabsorbed in the distal tubule. Most of the remaining 5 percent is reabsorbed in the collecting tubule. The amount of fluid, at this point called urine, that reaches the pelvis of the kidney is only 1 percent of the volume originally filtered at the glomerulus; but it contains nearly all the nitrogenous waste of the filtrate in concentrated solution. A few substances are also secreted from the blood through the walls of the tubule into the tubular fluid.

The action of the loop of Henle is more difficult to describe, and a full account is given below under *The human excretory system.*

Characteristics of the uric acid of birds and reptiles

Birds and reptiles. The main excretory product of birds and reptiles is uric acid. Since their glomeruli are relatively small, so also is their daily volume of urine. Not highly concentrated by mammalian standards—although it may be turbid with crystals of uric acid—the urine of birds and reptiles is conducted not to a urinary bladder but to the terminal portion of the alimentary canal, the cloaca; from the cloaca it is voided with the feces. Like mammals, and unlike the lower vertebrates, birds and reptiles have skins impermeable to water and thus are well adapted to terrestrial life. The relative inability of the kidney to produce concentrated urine is compensated for in birds that possess salt glands, which remove excess salt from their bodies. These organs are modified tear glands that discharge a concentrated solution of sodium chloride through the nostrils. Salt glands enable marine birds to drink seawater with no ill effects.

Amphibians. Direct evidence for the occurrence of filtration at the glomerulus was first provided by experiments on the amphibian kidney. Although amphibians are formally given the status of terrestrial animals, they are poorly adapted to life on land. They excrete nitrogen in the form of urea and cannot produce urine more concentrated than the blood. Their skins are permeable to water. On land amphibians are liable to lose water very rapidly by evaporation. In fresh water they suffer entry of water by osmosis, which is counteracted by the excretion of a large volume of dilute urine. The urine is stored in a large bladder before being voided, providing a reserve of water the animal can use when it comes on land.

When an amphibian leaves the water, a number of physiological adjustments are made that have the effect of conserving water. The rate of glomerular filtration is reduced by restriction of the blood supply, and this together with an increased release of antidiuretic hormone results in the production of a small volume of urine of the same concentration as the blood. Antidiuretic hormone (ADH, also known as vasopressin, which increases the permeability of the distal and collecting tubules to water) also increases the permeability of the bladder to water and allows the stored urine to be reabsorbed into the body.

Fishes. The homeostasis problem is the same for freshwater fishes as for other freshwater animals. Water enters the body by osmosis and salts leach out. To compensate, the kidney (which has large glomeruli) produces a relatively large amount of dilute urine (about 20 percent of the body weight per day). This serves to remove the water but by itself is insufficient to prevent gradual loss of salts. Extremely diluted salts are taken up from the fresh water and transported directly into the blood by certain specialized cells in the gills. Nitrogenous excretion is no problem: some ammonia is carried away in the large volume of dilute urine, but most of it simply escapes to the external medium by diffusing through the gills.

Excretion in marine fishes

By contrast, the homeostasis problem of marine fishes is unlike that of most marine animals. The salt content of the blood of marine fishes is less than half that of seawater (see below *Evolution of the vertebrate excretory system*); consequently, marine fishes tend to lose water and gain salt. This, it would seem, could be compensated most easily by the excretion of urine more concentrated than the blood, but the kidneys of fishes are not able to do this. In marine bony fishes the kidney has small glomeruli and produces only a small amount (about 4 percent of the body weight per day) of urine, which is of the same concentration as the blood. The fish replaces its lost water by continually swallowing seawater, and the special cells of the gills, working in reverse, reject salt to the external medium. Nitrogen is excreted mostly as ammonia but also as another detoxication product, trimethylamine oxide.

In sharks and rays ammonia is converted to urea, and urea plays an important role in homeostasis. Urea is retained in the blood to such an extent that the blood is slightly more concentrated than seawater. Thus loss of water by osmosis is prevented and these fish have no need to swallow seawater. Any excess of salt in their bodies is removed via the rectal gland, functionally analogous to the salt gland of birds.

Osmotic and ionic regulation in fishes is under hormonal control. This has been studied particularly in fishes such as eels and salmon, which are able to move between fresh water and seawater.

Evolution of the vertebrate excretory system. Studies of the embryonic development of primitive vertebrates, such as the dogfish shark, clearly show that the excretory system arises from a series of tubules, one pair in every segment of the body between the heart and the tail. This continuous series of tubules constitutes the archinephros, the name implying that the kidney of the ancestral vertebrate had some such form as this. Each tubule opens internally to the body cavity and may, in the remote past, have opened separately to the exterior; but in all living vertebrates the tubules open on each side into a longitudinal duct, the archinephric duct. At the posterior end of the body cavity the two archinephric ducts unite before opening to the exterior. Later in development, Bowman's capsule arises as a diverticulum of each tubule, subsequently becoming indented by the glomerulus. Eventually, the tubules usually lose their internal openings to the body cavity. The most anterior tubules of the archinephros (pronephros) usually degenerate in the adult.

Relationship between excretory and reproductive systems

These ducts and tubules also subserve the reproductive function, and for this reason they are also called the urogenital system. The extent to which the ducts and tubules are shared is greater in the male than in the female. In the male the spermatic tubules of the testis connect with the kidney tubules in the middle region of the archinephros (mesonephros), and in some vertebrates (*e.g.,* the frog) where there is no development of the posterior region (metanephros), the tubules of the mesonephros serve to convey both urine and sperm. In the reptiles, birds, and mammals there is greater separation of function, the mesonephros being exclusively genital and the metanephros being exclusively urinary (see Figure 3).

In the female, even in the lower vertebrates, the two systems are confluent only at the posterior end. It has been held that the oviduct is a derivative of the archinephric duct, but the evidence for this is not compelling.

In primitive marine animals the blood is almost identical with seawater in composition; in typical freshwater animals the concentration of the blood is about half that of seawater. Many originally marine animals have evolved the ability to live in fresh water; relatively few animals, after having thus evolved, have returned to the sea, and in none of them has the blood returned to its original "seawa-

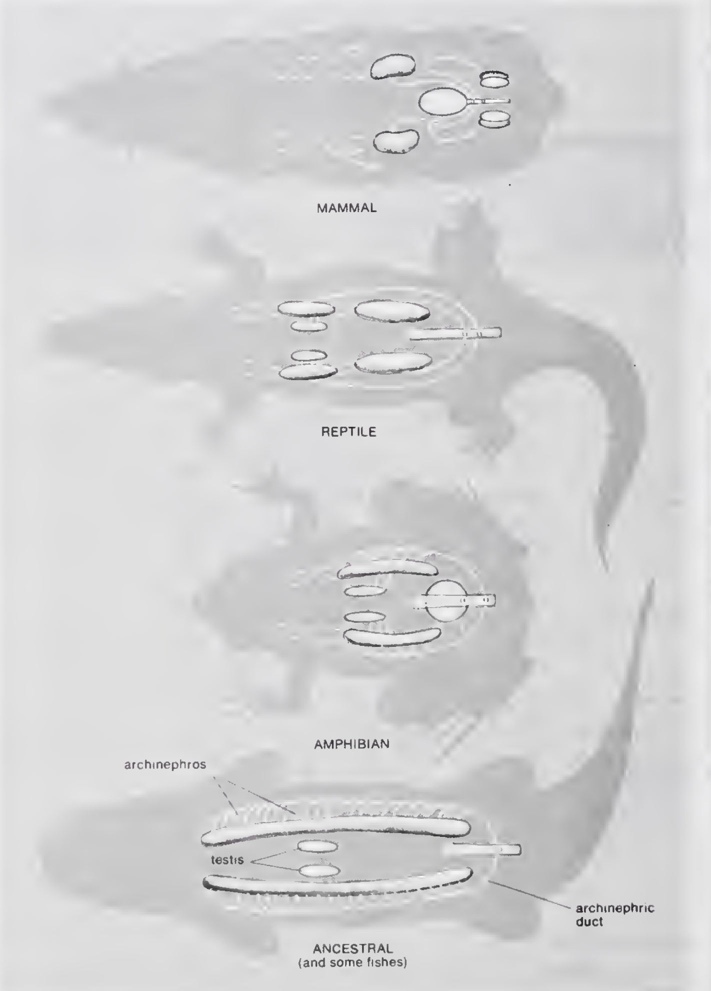


Figure 3: *Evolution of the urogenital system in male vertebrates.* The ancestral archinephric condition has evolved in different patterns for the amphibians, reptiles, and mammals (birds have a pattern similar to reptiles). The broken lines indicate the change in ancestral plan in each of the major vertebrate classes.

ter" concentration. The earliest fossil vertebrates are found in marine deposits, but the fossil record shows clearly that the early evolution of fishes took place in fresh water. It is assumed that the blood of early freshwater fishes, like that of other freshwater animals, was osmotically equivalent to half-strength seawater. The sharks and rays returned to the sea during the Carboniferous Period, and no doubt at that time they evolved the device of urea retention. The bony fishes returned to the sea later, in the Mesozoic Era, and solved their problem by swallowing seawater and rejecting excess salt at the gills. (J.A.R./Ed.)

The human excretory system

In many respects the human excretory, or urinary, system resembles those of other mammalian species, but it has its own unique structural and functional characteristics. The terms excretory and urinary emphasize the eliminatory function of the system. The kidneys, however, both secrete and actively retain within the body certain substances that are as critical to survival as those that are eliminated.

The system contains two kidneys, which control the electrolyte composition of the blood and eliminate dissolved waste products and excess amounts of other substances from the blood; the latter substances are excreted in the urine, which passes from the kidneys to the bladder by way of two thin muscular tubes called the ureters. The bladder is a sac that holds the urine until it is eliminated through the urethra (Figure 4).

THE KIDNEYS

General description and location. The kidneys are bean-shaped, reddish brown paired organs, concave on one long side and convex on the opposite. They are normally located high in the abdominal cavity and against its back wall, lying on either side of the vertebral column between the levels of the 12th thoracic and third lumbar vertebrae, and outside the peritoneum, the membrane that lines the abdomen.

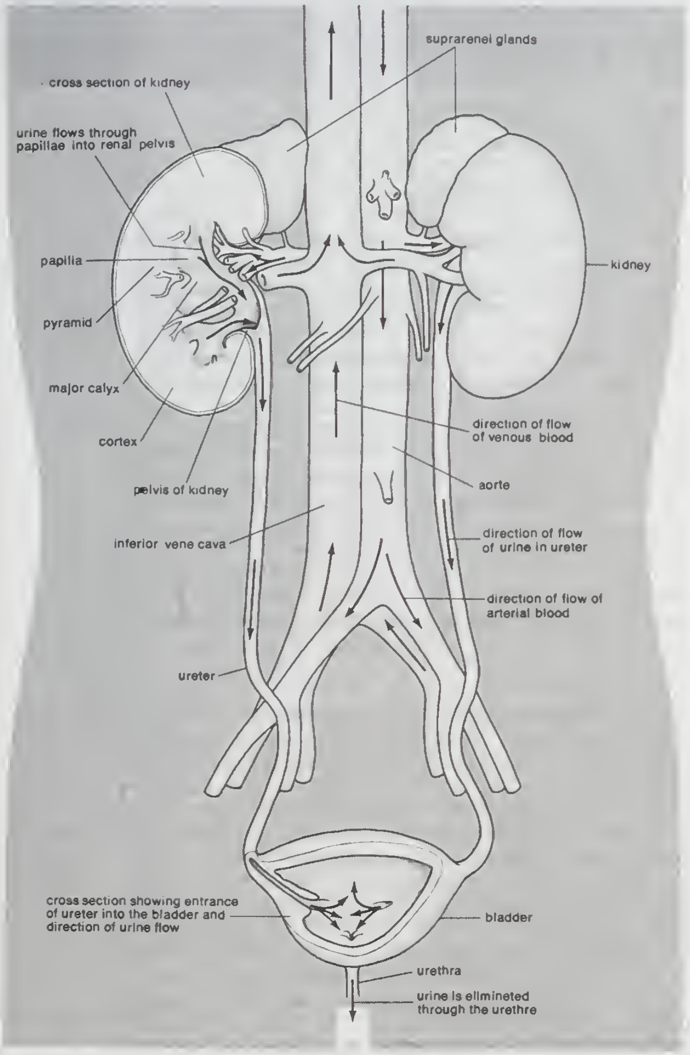

suprarenal glands
cross section of kidney
urine flows through papillae into renal pelvis
papilla
kidney
pyramid
major calyx
cortex
direction of flow of venous blood
pelvis of kidney
aorta
inferior vena cava
direction of flow of urine in ureter
direction of flow of arterial blood
ureter
cross section showing entrance of ureter into the bladder and direction of urine flow
bladder
urethra
urine is eliminated through the urethra

Figure 4: Human renal excretory system.

The long axes of the kidneys are aligned with that of the body, but the upper end of each kidney (pole) is tilted slightly inward toward the backbone (vertebral column). Situated in the middle of the medial concave border is a deep vertical cleft, the hilus, which leads to a cavity within the kidney known as the renal (kidney) sinus. The hilus is the point of entry and exit of the renal arteries and veins, lymphatic vessels, nerves, and the enlarged upper extension of the ureters.

Renal vessels and nerves. The renal arteries arise, one on each side, from the abdominal aorta at a point opposite the upper border of the second lumbar vertebra (*i.e.*, a little above the small of the back). Close to the renal hilus each artery gives off small branches to the adrenal gland and ureter and then branches into anterior and posterior divisions. The large veins carrying blood from the kidneys usually lie in front of the corresponding arteries and join the inferior vena cava almost at right angles. The left vein is longer than the right vein because the inferior vena cava lies closer to the right kidney.

The kidneys are supplied with sympathetic and parasympathetic nerves of the autonomic nervous system, and the renal nerves contain both afferent and efferent fibres (afferent fibres carry nerve impulses to the central nervous system; efferent fibres, from it).

Types of kidney tissue

Internal configuration. A cross section of a kidney reveals the renal sinus and two layers of kidney tissue distinguishable by their texture and colour. The innermost tissue, called the renal medulla, forms comparatively dark cones, called renal pyramids, with bases outward and apexes projecting, either singly or in groups, into the renal sinus. Each projection of one or more pyramid apexes into the sinus is known as a renal papilla. The bases of these pyramids are irregular, with slender striations extending toward the external kidney surface. The paler, more granular tissue external to the medulla is the cortex. It arches over the bases of the pyramids and fills gaps between the pyramids. Each group of pyramids that projects into a papilla, together with the portion of cortex that arches over the group, is called a renal lobe.

The renal sinus includes the renal pelvis, a funnel-shaped expansion of the upper end of the ureter, and, reaching into the kidney substances from the wide end of the funnel, two or three extensions of the cavity called the major calyxes. The major calyxes are divided in turn into four to 12 smaller cuplike cavities, the minor calyxes, into which the renal papillae project. The renal pelvis serves as the initial reservoir for urine, which flows into the sinus through the urinary collecting tubules, small tubes that open into the sinus at the papillae.

The nephrons

Minute structure. The structural units of the kidneys that actually produce urine are the nephrons, of which there are approximately 1,000,000 in each kidney. Each nephron is a long tubule (or extremely fine tube) that is closed, expanded, and folded into a double-walled cuplike structure at one end. This structure, called the renal corpuscular capsule, or Bowman's capsule, encloses a cluster of capillaries (microscopic blood vessels) called the glomerulus. The capsule and glomerulus together constitute a renal corpuscle, also called a malpighian body. Blood flows into and away from the glomerulus through small arteries (arterioles) that enter and exit the glomerulus through the open end of the capsule. This opening is called the vascular pole of the corpuscle.

The tubules of the nephrons are 30–55 millimetres (1.2–2.2 inches) long. The corpuscle and the initial portion of each tubule, called the proximal convoluted tubule, lie in the renal cortex. The tubule descends into a renal pyramid, makes a U-shaped turn, and returns to the cortex at a point near its point of entry into the medulla. This section of the tubule, consisting of the two parallel lengths and the bend between them, is called the loop of Henle or the nephronic loop. After its reentrance into the cortex, the tubule returns to the vascular pole (the opening in the cuplike structure of the capsule) of its own nephron. The final portion of the tubule, the distal convoluted tubule, leads from the vascular pole of the corpuscle to a collecting tubule, by way of a short junctional tubule. Several of the collecting tubules join together to form a somewhat wider tubule, which carries the urine to a renal papilla and the renal pelvis.

Although all nephrons in the kidney have the same general disposition, there are regional differences, particularly in the length of the loops of Henle. Glomeruli that lie deep in the renal cortex near the medulla (juxtamedullary glomeruli) possess long loops of Henle that pass deeply into the medulla, whereas more superficial cortical glomeruli have much shorter loops. Among different animal species the length of the loops varies considerably and affects the ability of the species to concentrate urine above the osmotic concentration of plasma.

The successive sections of the nephron tubule vary in shape and calibre, and these differences, together with differences in the cells that line the sections, are associated with specific functions in the production of urine.

Intrarenal network of blood vessels. The intrarenal network of blood vessels forms part of the blood-processing apparatus of the kidneys.

Arteries and arterioles. The anterior and posterior divisions of each renal artery, mentioned earlier, divide into lobar arteries, each of which enters the kidney substance through or near a renal papilla. Each lobar artery gives off two or three branches, called interlobar arteries, which run outward between adjacent renal pyramids. When these reach the boundary between the cortex and the medulla they split almost at right angles into branches called arcuate arteries that curve along between the cortex and the medulla parallel to the surface of the kidney. Many arteries, called interlobular arteries, branch off from the arcuate arteries and radiate out through the cortex to end in networks of capillaries in the region just inside the capsule. En route they give off short branches called the afferent arterioles, which carry blood to the glomeruli where they divide into four to eight loops of capillaries in each glomerulus (Figure 5).

Near and before the point where the afferent arteriole enters the glomerulus, its lining layer becomes enlarged and contains secretory granules. This composite structure is called the juxtaglomerular apparatus (JGA) and is believed to be involved in the secretion of renin (see below *The role of hormones in renal function*). They are then reconstituted near the point of entry of the afferent arteriole to become the efferent arterioles carrying blood away from the glomeruli. The afferent arterioles are almost twice as thick as the efferent arterioles because they have thicker muscular coats, but the sizes of their channels are almost the same.

Throughout most of the cortex the efferent arterioles redivide into a second set of capillaries, which supply blood to the proximal and distal renal tubules.

Arterioles in the juxtamedullary zone

The efferent glomerular arterioles of juxtaglomerular glomeruli divide into vessels that supply the contiguous tubules and vessels that enter the bases of the renal pyramids. Known as vasa recta, these vessels run toward the apexes of the pyramids in close contact with the loops of Henle. Like the tubules they make hairpin bends, retrace their path, and empty into arcuate veins that parallel the arcuate arteries.

Normally the blood circulating in the cortex is more abundant than that in the medulla (amounting to over 90 percent of the total), but in certain conditions, such as those associated with severe trauma or blood loss, cortical vessels may become constricted while the juxtamedullary circulation is preserved. Because the cortical glomeruli and tubules are deprived of blood, the flow of urine is diminished, and in extreme cases may cease.

Veins and venules. The renal venules (small veins) and veins accompany the arterioles and arteries and are referred to by similar names. The venules that lie just beneath the renal capsule, called stellate venules because of their radial arrangement, drain into interlobular venules. In turn these combine to form the tributaries of the arcuate, interlobar, and lobar veins. Blood from the renal pyramids passes into vessels, called venae rectae, which join the arcuate veins. In the renal sinus the lobar veins unite to form veins corresponding to the main divisions of the renal arteries, and they normally fuse to constitute a single renal vein in or near the renal hilus.

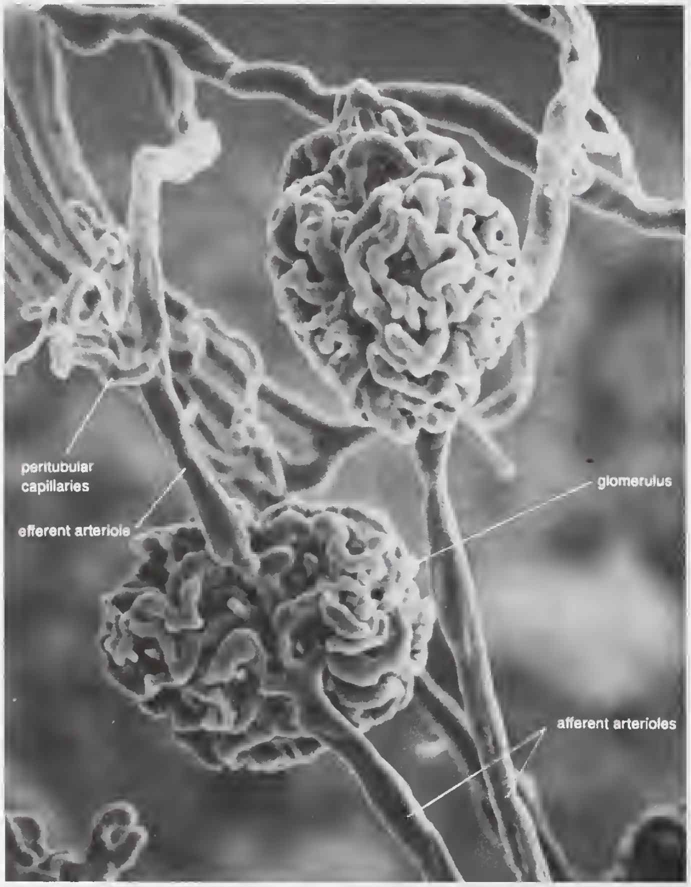


Figure 5: A cast of the intrarenal vascular network. The afferent arteriole enters the glomerulus, where it divides into capillary loops. These are rejoined to become the efferent arteriole, which leaves the glomerulus and branches into numerous peritubular capillaries.

By courtesy of Andrew P. Evan and Vincent H. Gattone II

Lymphatic network. Lymphatic capillaries form a network just inside the renal capsule and another, deeper network between and around the renal blood vessels. Few lymphatic capillaries appear in the actual renal substance, and those present are evidently associated with the connective tissue framework, while the glomeruli contain no lymphatics. The lymphatic networks inside the capsule and around the renal blood vessels drain into lymphatic channels accompanying the interlobular and arcuate blood vessels. The main lymph channels run alongside the main renal arteries and veins to end in lymph nodes beside the aorta and near the sites of origin of the renal arteries.

THE URETERS

General characteristics. The ureters are narrow, thick-walled ducts, about 25–30 centimetres (9.8–11.8 inches) in length and from four to five millimetres (0.16 to 0.2 inch) in diameter, that transport the urine from the kidneys to the urinary bladder. Throughout their course they lie behind the peritoneum, the lining of the abdomen and pelvis, and are attached to it by connective tissue.

In both sexes the ureters enter the bladder wall about five centimetres apart, although this distance is increased when the bladder is distended with urine. The ureters run obliquely through the muscular wall of the bladder for nearly two centimetres before opening into the bladder cavity through narrow apertures. This oblique course provides a kind of valvular mechanism; when the bladder becomes distended it presses against the part of each ureter that is in the muscular wall of the bladder, and this helps to prevent the flow of urine back into the ureters from the bladder.

Structure of the ureteric wall. The wall of the ureter has three layers, the adventitia, or outer layer; the intermediate, muscular layer; and the lining, made up of mucous membrane. The adventitia consists of fibroelastic connective tissue that merges with the connective tissue behind the peritoneum. The muscular coat is composed of smooth (involuntary) muscle fibres and, in the upper two-thirds of the ureter, has two layers—an inner layer of fibres arranged longitudinally and an outer layer disposed circularly. In the lower third of the ureter an additional longitudinal layer appears on the outside of the vessel. As each ureter extends into the bladder wall its circular fibres disappear, but its longitudinal fibres extend almost as far as the mucous membrane lining the bladder.

The mucous membrane lining increases in thickness from the renal pelvis downward. Thus, in the pelvis and the calyxes of the kidney the lining is two to three cells deep; in the ureter, four to five cells thick; and in the bladder, six to eight cells. The mucous membrane of the ureters is arranged in longitudinal folds, permitting considerable dilation of the channel. There are no true glands in the mucous membrane of the ureter or of the renal pelvis. The chief propelling force for the passage of urine from the kidney to the bladder is produced by peristaltic (wavelike) movements in the ureter muscles.

Peristalsis in the ureters

THE URINARY BLADDER

General description. The urinary bladder is a hollow muscular organ forming the main urinary reservoir. It rests on the anterior part of the pelvic floor (see below), behind the symphysis pubis and below the peritoneum. (The symphysis pubis is the joint in the hip bones in the front midline of the body.) The shape and size of the bladder vary according to the amount of urine that the organ contains. When empty it is tetrahedral and lies within the pelvis; when distended it becomes ovoid and expands into the lower abdomen. It has a body, with a fundus, or base; a neck; an apex; and a superior (upper) and two inferolateral (below and to the side) surfaces, although these features are not clearly evident except when the bladder is empty or only slightly distended.

The neck of the bladder is the area immediately surrounding the urethral opening; it is the lowest and most fixed part of the organ. In the male it is firmly attached to the base of the prostate, a gland that encircles the urethra.

The superior surface of the bladder is triangular and is covered with peritoneum. The bladder is supported on the levator ani muscles, which constitute the major part of the floor of the pelvic cavity. The bladder is covered, and to a certain extent supported, by the visceral layer of the pelvic fascia. This fascial layer is a sheet of connective tissue that sheaths the organs, blood vessels, and nerves of the pelvic cavity. The fascia forms, in front and to the side, ligaments, called pubovesical ligaments, that act as a kind of hammock under the inferolateral surfaces and neck of the bladder.

Blood and nerve supplies. The blood supply of the bladder is derived from the superior, middle, and inferior vesical (bladder) arteries. The superior vesical artery supplies the dome of the bladder, and one of its branches (in males) gives off the artery to the ductus deferens, a part of the passageway for sperm. The middle vesical artery supplies the base of the bladder. The inferior vesical artery supplies the inferolateral surfaces of the bladder and assists in supplying the base of the bladder, the lower end of the ureter, and other adjacent structures.

The nerves to the bladder

The nerves to the urinary bladder belong to the sympathetic and the parasympathetic divisions of the autonomic nervous system. The sympathetic nerve fibres come from the hypogastric plexus of nerves that lie in front of the fifth lumbar vertebra. Sympathetic nerves carry to the central nervous system the sensations associated with distention of the bladder and are believed to be involved in relaxation of the muscular layer of the vesical wall and with contraction of sphincter mechanism that closes the opening into the urethra. The parasympathetic nerves travel to the bladder with pelvic splanchnic nerves from the second through fifth sacral spinal segment. Parasympathetic nerves are concerned with contraction of the muscular walls of the bladder and with relaxation of its sphincter. Consequently they are actively involved in urination and are sometimes referred to as the emptying, or detrusor, nerves.

Structure of the bladder wall. The bladder wall has a serous coat over its upper surface. This covering is a continuation of the peritoneum that lines the abdominal cavity; it is called serous because it exudes a slight amount of lubricating fluid called serum. The other layers of the bladder wall are the fascial, muscular, submucous, and mucous coats.

The fascial coat is a layer of connective tissue, such as that which covers muscles. The muscular coat consists of coarse fascicles, or bundles, of smooth (involuntary) muscle fibres arranged in three strata, with fibres of the outer and inner layers running lengthwise, and with fibres of the intermediate layer running circularly; there is considerable intermingling of fibres between the layers. The smooth muscle coat constitutes the powerful detrusor muscle, which causes the bladder to empty.

The circular or intermediate muscular stratum of the vesical wall is thicker than the other layers. Its fibres, although running in a generally circular direction, do interlace. The internal muscular stratum is an indefinite layer of fibres that are mostly directed longitudinally. The submucous coat consists of loose connective tissue containing many elastic fibres. It is absent in the trigone, a triangular area whose angles are at the two openings for the ureters and the single internal urethral opening. Slim bands of muscle run between each ureteric opening and the internal urethral orifice; these are thought to maintain the oblique direction of the ureters during contraction of the bladder. Another bundle of muscle fibres connects the two ureteric openings and produces a slightly downwardly curved fold of mucous membrane between the openings.

The mucous coat of the bladder

The mucous coat, the innermost lining of the bladder, is an elastic layer impervious to urine. Over the trigone it firmly adheres to the muscular coat and is always smooth and pink whether the bladder is contracted or distended. Elsewhere, if the bladder is contracted, the mucous coat has multiple folds and a red, velvety appearance. When the bladder is distended, the folds are obliterated, but the difference in colour between the paler trigonal area and the other areas of the mucous membrane persists. The mucous membrane lining the bladder is continuous with that lining the ureters and the urethra.

THE URETHRA

The male urethra

General description. The urethra is the channel that conveys the urine from the bladder to the exterior. In the male it is about 20 centimetres long and carries not only the urine but also the semen and the secretions of the prostate, bulbourethral, and urethral glands. During urination and ejaculation it opens up, and its diameter then varies from 0.5 to 0.8 centimetre along its length, but at other times its walls touch and its lining is raised into longitudinal folds. The male urethra has three distinguishable parts, the prostatic, the membranous, and the spongy, each part being named from the structures through which it passes rather than from any inherent characteristics.

The prostatic section of the male urethra commences at the internal urethral orifice and descends almost vertically through the prostate, from the base of the gland to the apex, describing a slight curve with its concavity forward. It is about 2.5 to three centimetres long and is spindle-shaped; its middle portion is the widest and most dilatable part of the urethra. The membranous part of the male urethra is in the area between the two layers of a membrane called the urogenital diaphragm. The urethra is narrower in this area than at any other point except at its external opening and is encircled by a muscle, the sphincter urethrae. The two small bulbourethral glands are on either side of it. The membranous urethra is not firmly attached to the layers of the urogenital diaphragm. The spongy part of the male urethra is that part of the urethra that traverses the penis. It passes through the corpus spongiosum of the penis. The ducts of the bulbourethral glands enter the spongy urethra about 2.5 centimetres below the lower layer of the urogenital membrane; except near its outer end, many mucous glands also open into it.

The female urethra

The female urethra is much shorter (three to 4.5 centimetres) and more distensible than the corresponding channel in males and carries only urine and the secretions of mucous glands. It begins at the internal opening of the urethra into the bladder and curves gently downward and forward through the urogenital diaphragm, where it is surrounded, as in the male, by the sphincter urethrae. It lies behind and below the symphysis pubis. Except for its uppermost part, the urethra is embedded in the anterior wall of the vagina. The external urethral orifice is immediately in front of the vaginal opening, about 2.5 centimetres behind the clitoris, and between the labia minora, the inner folds at the outer opening of the vagina.

Structure of urethral wall. The urethra of the male is a tube of mucous membrane supported on a submucous layer and an incomplete muscular coat. The membrane forms longitudinal folds when the tube is empty; these folds are more prominent in the membranous and spongy parts. There are many glands in the mucous membrane, and they are more common in the posterior wall of the spongy part. The submucous layer is composed of fibroelastic connective tissue containing numerous small blood vessels, including more venules than arterioles. The thin muscular coat consists of smooth (involuntary) and striated (voluntary) muscle fibres. The smooth muscular layer, longitudinally disposed, is continuous above with the detrusor muscle of the bladder and extends distally as far as the membranous urethra, where it is replaced and partly surrounded by striated muscle of the external sphincter. The somatic nerves to the external sphincter are the efferent and afferent components of the pudendal nerve, arising from the second, third, and fourth sacral segments of the spinal cord.

The female urethra has mucous, submucous, and muscular coats. As in the male, the lining of the empty channel is raised into longitudinal folds. It also shows mucous glands, mentioned in the preceding paragraphs as existing in the male urethra. The submucous coat resembles that in the male, except that the venules are even more prominent. In both sexes, but especially in females, this layer appears to be a variety of erectile tissue. The muscular coat extends along the entire length of the female urethra and is continuous above with the musculature of the bladder. It consists of inner longitudinal and outer circular layers, and fibres from the latter intermix with those in the anterior wall of the vagina, in which the urethra is embedded.

(G.A.G.M./J.S.Ro.)

Human excretion

GENERAL FUNCTION OF THE KIDNEY

The kidney has evolved so as to enable humans to exist on land where water and salts must be conserved, wastes excreted in concentrated form, and the blood and the tissue fluids strictly regulated as to volume, chemical composition, and osmotic pressure. Under the drive of arterial pressure, water and salts are filtered from the blood through the capillaries of the glomerulus into the lumen, or passageway, of the nephron, and then most of the water and the substances that are essential to the body are reabsorbed into the blood. The remaining filtrate is drained off as urine. The kidneys, thus, help maintain a constant internal environment despite a wide range of changes in the external environment.

Regulatory functions. The kidneys regulate three essential and interrelated properties of the tissues—water content, acid-base balance, and osmotic pressure—in such a way as to maintain electrolyte and water equilibrium; in other words, the kidneys are able to maintain a balance between quantities of water and the quantities of such chemicals as calcium, potassium, sodium, phosphorus, and sulfate in solution. Unless the concentrations of mineral ions such as sodium, crystalloids such as glucose, and wastes such as urea are maintained within narrow normal limits, bodily malfunction rapidly develops leading to sickness or death.

The removal of both kidneys causes urinary constituents to accumulate in the blood (uremia), resulting in death in 14–21 days if untreated. (The term uremia does not mean that urea is itself a toxic compound responsible for illness and death.) Whenever the blood contains an abnormal constituent in solution or an excess of normal constituents including water and salts, the kidneys excrete these until normal composition is restored. The kidneys are the only means for eliminating the wastes that are the end products of protein metabolism. They do not themselves modify the waste products that they excrete, but transfer them to the urine in the form in which they are produced in other parts of the body. The only exception to this is their ability to manufacture ammonia. The kidneys also eliminate drugs and toxic agents. Thus, the kidneys eliminate the unwanted end products of metabolism, such as urea, while limiting the loss of valuable substances, such as glucose. In maintaining the acid-base equilibrium, the kidneys remove the excess of hydrogen ions produced from the normally acid-forming diet and manufacture ammonia to remove these ions in the urine as ammonium salts.

Volumes of blood, water, and urine processed by kidneys

To carry on its functions the kidney is endowed with a relatively huge blood supply. The blood processed in the kidneys amounts to some 1,200 millilitres a minute, or 1,800 litres (about 475 gallons) a day, which is 400 times the total blood volume and roughly one-fourth the volume pumped each day by the heart. Every 24 hours 170 litres (45 gallons) of water are filtered from the bloodstream into the renal tubules; and by far the greater part of this—some 168.5 litres of water together with salts dissolved in it—is reabsorbed by the cells lining the tubules and returned to the blood. The total glomerular filtrate in 24 hours is no less than 50–60 times the volume of blood plasma (the blood minus its cells) in the entire body. In a 24-hour period, an average man eliminates only 1.5 litres of water, containing the waste products of metabolism, but the actual volume varies with fluid intake and occupational and environmental factors. With vigorous sweating it may fall to 500 millilitres (about a pint) a day; with a large water intake it may rise to three litres, or six times as much. The kidney can vary its reabsorption of water to compensate for changes in plasma volume resulting from dehydration or overhydration.

Nonexcretory functions. The kidneys also perform certain nonexcretory functions. They secrete substances that enter the blood. These are of three kinds: renin, which is concerned indirectly with the control of electrolyte balance and blood pressure; erythropoietin, which is important for the formation of hemoglobin and red blood cells, especially in response to anemia or deficiency of oxygen reaching the body tissues; and 1,25-dihydroxycholecalciferol, which is the metabolically active form of vitamin D. Finally, although the kidneys are subject to both nervous and humoral (hormonal) control, they do possess a considerable degree of autonomy; *i.e.,* function continues in an organ isolated from the nervous system but kept alive with circulating fluid. Indeed, if this were not so kidney transplantation would be impossible.

RENAL BLOOD CIRCULATION

Intrarenal blood pressures. The renal arteries are short and spring directly from the abdominal aorta, so that arterial blood is delivered to the kidneys at maximum available pressure. As in other vascular beds, renal perfusion is determined by the renal arterial blood pressure and vascular resistance to blood flow. Evidence indicates that in the kidneys the greater part of the total resistance occurs in the glomerular arterioles. The muscular coats of the arterioles are well supplied with sympathetic vasoconstrictor fibres (nerve fibres that induce narrowing of the blood vessels), and there is also a small parasympathetic supply from the vagus and splanchnic nerves that induces dilation of the vessels. Sympathetic stimulation causes vasoconstriction and reduces urinary output. The vessel walls are also sensitive to circulating epinephrine and norepinephrine hormones, small amounts of which constrict the efferent arterioles and large amounts of which constrict all the vessels; and to angiotensin, which is a constrictor agent closely related to renin. Prostaglandins may also have a role.

Factors that affect renal flow. The kidney is able to regulate its internal circulation regardless of the systemic blood pressure, provided that the latter is not extremely high or extremely low. The forces that are involved in maintaining a circulation of the blood in the kidneys must remain constant if the monitoring of the water and electrolyte composition of the blood is to proceed undisturbed. This regulation is preserved even in the kidney cut off from the nervous system and, to a lesser extent, in an organ removed from the body and kept viable by having salt solutions of physiologically suitable concentrations circulated through it; it is commonly referred to as autoregulation.

The exact mechanism by which the kidney regulates its own circulation is not known, but various theories have been proposed: (1) Smooth muscle cells in the arterioles may have an intrinsic basal tone (normal degree of contraction) when not affected by nervous or humoral (hormonal) stimuli. The tone responds to alterations in perfusion pressure in such a way that when the pressure falls the degree of contraction is reduced, preglomerular resistance is lowered, and blood flow is preserved. Conversely, when perfusion pressure rises, the degree of contraction is increased and blood flow remains constant. (2) If the renal blood flow rises, more sodium is present in the fluid in the distal tubules because the filtration rate increases. This rise in the sodium level stimulates the secretion of renin from the JGA with the formation of angiotensin, causing the arterioles to constrict and blood flow to be reduced. (3) If systemic blood pressure rises, the renal blood flow remains constant because of the increased viscosity of the blood. Normally, the interlobular arteries have an axial (central) stream of red blood cells with an outer layer of plasma so that the afferent arterioles skim off more plasma than cells. If the arteriolar blood pressure rises, the skimming effect increases, and the more densely packed axial flow of cells in the vessels offers increasing resistance to the pressure, which has to overcome this heightened viscosity. Thus, the overall renal blood flow changes little. Up to a point, similar considerations in reverse apply to the effects of reduced systemic pressure. (4) Changes in the arterial pressure modify the pressure exerted by the interstitial (tissue) fluid of the kidney on capillaries and veins so that increased pressure raises, and decreased pressure lowers, resistance to blood flow.

The renal blood flow is greater when a person is lying down than when standing; it is higher in fever; and it is reduced by prolonged vigorous exertion, pain, anxiety, and other emotions that constrict the arterioles and divert blood to other organs. It is also reduced by hemorrhage and asphyxia and by depletion of water and salts, which is severe in shock, including operative shock. A large fall in systemic blood pressure, as after severe hemorrhage, may so reduce renal blood flow that no urine at all is formed for a time; death may occur from suppression of glomerular function. Simple fainting causes vasoconstriction and reduced urine output. Urinary secretion is also stopped by obstruction of the ureter when back pressure reaches a critical point.

Glomerular pressure. The importance of these various vascular factors lies in the fact that the basic process occurring in the glomerulus is one of filtration, the energy for which is furnished by the blood pressure within the glomerular capillaries. Glomerular pressure is a function of the systemic pressure as modified by the tone (state of constriction or dilation) of the afferent and efferent arterioles, as these open or close spontaneously or in response to nervous or hormonal control.

In normal circumstances glomerular pressure is believed to be about 45 millimetres of mercury (mmHg), which is a higher pressure than that found in capillaries elsewhere in the body. As is the case in renal blood flow, the glomerular filtration rate is also kept within the limits between which autoregulation of blood flow operates. Outside these limits, however, major changes in blood flow occur. Thus, severe constriction of the afferent vessels reduces blood flow, glomerular pressure, and filtration rate, while efferent constriction causes reduced blood flow but increases glomerular pressure and filtration.

FORMATION AND COMPOSITION OF URINE

The urine leaving the kidney differs considerably in composition from the plasma entering it (Table 1). The study of renal function must account for these differences; *e.g.,* the absence of protein and glucose from the urine, a change in the pH of urine as compared with that of plasma, and the high levels of ammonia and creatinine in the urine, while sodium and calcium remain at similar low levels in both urine and plasma.

Table 1: Relative Composition of Plasma and Urine in Normal Men

	plasma g/100 ml	urine g/100 ml	concentration in urine
Water	90–93	95	—
Protein	7–8.5	—	—
Urea	0.03	2	× 60
Uric acid	0.002	0.03	× 15
Glucose	0.1	—	—
Creatinine	0.001	0.1	× 100
Sodium	0.32	0.6	× 2
Potassium	0.02	0.15	× 7
Calcium	0.01	0.015	× 1.5
Magnesium	0.0025	0.01	× 4
Chloride	0.37	0.6	× 2
Phosphate	0.003	0.12	× 40
Sulfate	0.003	0.18	× 60
Ammonia	0.0001	0.05	× 500

A large volume of ultrafiltrate (*i.e.,* a liquid from which the blood cells and the blood proteins have been filtered out) is produced by the glomerulus into the capsule. As this liquid traverses the proximal convoluted tubule, most of its water and salts are reabsorbed, some of the solutes completely and others partially; *i.e.,* there is a separation of substances that must be retained from those due for rejection. Subsequently the loop of Henle, distal convoluted tubule, and collecting ducts are mainly concerned with the fine control of water and electrolyte balance (Figure 6).

From J.M. Forrester, R. Passmore, and J.S. Robson (eds.), *A Companion to Medical Studies,* vol. 1, 3rd ed. (1985), Blackwell Scientific Publications Ltd

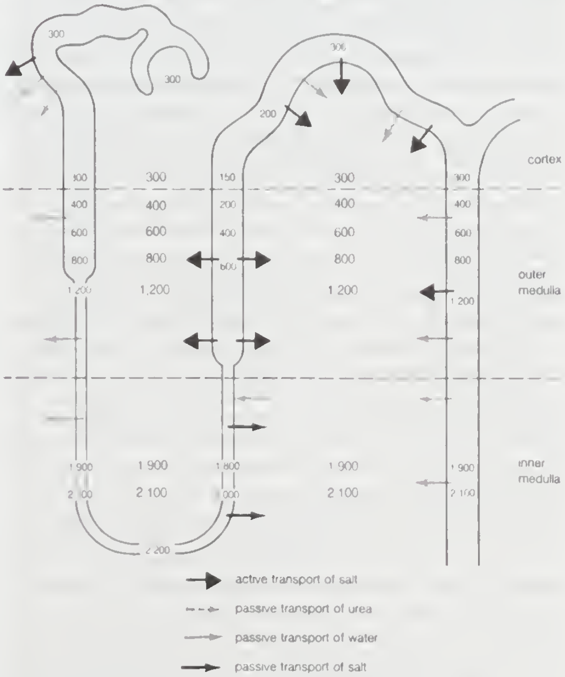


Figure 6: The modes of action of the countercurrent mechanism for concentrated urine production. The values indicated are given in mosmoles of solute/kilogram.

Glomerular filtration. Urine formation begins as a process of ultrafiltration of a large volume of blood plasma from the glomerular capillaries into the capsular space, colloids such as proteins being held back while crystalloids (substances in true solution) pass through. In humans, the average capillary diameter is five to 10 micrometres (a micrometre is 0.001 millimetre). The wall of each loop of capillaries has three layers (Figure 7). The inner layer

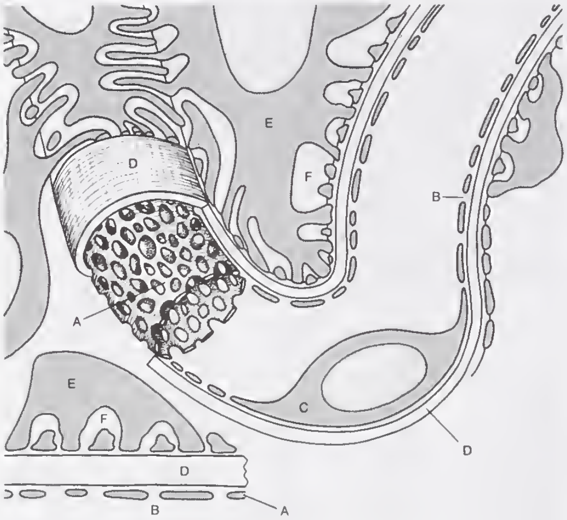


Figure 7: The structure of the human glomerulus, depicting (A) the fenestrated capillary endothelium; (B) fenestrae; (C) an endothelial cell; (D) the basement membrane; (E) a podocyte; and (F) a pedicel.

From J M Forrester, R. Passmore, and J S Robson (eds), *A Companion to Medical Studies*, vol 1, 3rd ed (1985), Blackwell Scientific Publications Ltd

consists of flat nucleated endothelial cells arranged to form numerous pores, or fenestrae, 50–100 nanometres in diameter (a nanometre is 0.000001 millimetre), which allow the blood to make direct contact with the second layer, a basement membrane. The basement membrane of the capillaries, similar to that which occurs in the lining of many other structures and organs, is a continuous layer of hydrated collagen and glycopeptides. Although once thought to be homogeneous, it appears to consist of three layers that differ in the content of polyanionic glycopeptides. The membrane is negatively charged (anionic), owing to its relatively high content of sialic and aspartic acids. Also present are glycosaminoglycans, such as heparin sulfate. The third, external layer consists of large epithelial cells called podocytes (Figure 8). These cells make contact with the outer surface of the basement membrane by slender cytoplasmic extensions called pedicels (foot processes). These processes are slightly expanded at their point of contact with the basement membrane and are separated from each other by slitlike spaces about 20 to 30 nanometres across. A fine membrane (slit diaphragm) closes the slitlike spaces near the basement membrane.

Passage of filtrate

There are two physical processes by which glomerular filtrate may pass the barrier of the glomerular wall—simple diffusion and bulk flow. In bulk flow, the solute in the glomerular filtrate with water passes through pores in the basement membrane. In either case the ultimate restriction to the passage of filtrate appears to lie in the hydrated gel structure of the basement membrane. The negative electrostatic charge in the membrane is an additional restrictive force for negatively charged anionic macromolecules, such as albumin (molecular weight 69,000), while larger protein molecules are restricted by size alone. On the other hand, proteins of smaller molecular size—*e.g.*, neutral gelatin (35,000)—pass through freely. It is possible that the endothelial cell layer may also help to exclude very large molecules and blood cells and that a similar effect is exerted by the slit pores and diaphragm.

The normal process of glomerular filtration depends upon the integrity of the glomerulus, which in turn depends upon its proper nutrition and oxygenation. If glomeruli are damaged through disease or lack of oxygen they become more permeable, allowing plasma proteins to enter the urine. Special cells that may be concerned with the formation and maintenance of the basement membrane of the glomerular walls are called mesangial cells. These lie between loops of the glomerular capillaries and form a stalk or scaffolding for the capillary network. They are themselves embedded in a matrix of glycosaminoglycan similar to that of the glomerular capillary basement membrane and may be responsible for its formation. The mesangial cells are also responsible for ridding the basement membrane of large foreign molecules that may be held there in the course of certain diseases. These cells proliferate and the mesangial matrix enlarges in the course of immunologically induced diseases affecting the glomerulus (see below *Glomerulonephritis*).

Tubule function. The role of the tubules may be assessed by comparing the amounts of various substances in the filtrate and in the urine (Table 2).

General function of tubules

It is apparent that the filtrate must be modified in the tubules to account for the differing compositions of filtrate and final urine; *e.g.*, to allow for the total absence of glucose in the latter, the much smaller volume of urine than filtrate, or for the acidity of urine compared with the neutrality of the filtrate.

As the filtrate passes along the proximal tubule, most of its water and salts are reabsorbed into the blood of the network of capillaries around the tubules. Of other substances, some are reabsorbed completely, others in part, because this portion of the nephron separates substances that must be retained in the body from those destined for excretion in the urine. The function of the proximal tubule is essentially reabsorption of filtrate in accordance with the needs of homeostasis (equilibrium), whereas the distal part of the nephron and collecting duct are mainly concerned with the detailed regulation of water, electrolyte, and hydrogen-ion balance. All of these processes occur in the tubules through both chemical and physical means, and all are subject to hormonal regulation. Although the urine normally differs markedly from filtrate, if tubule function is progressively reduced in experimental situations by cooling or poisoning, the urine will come increasingly to resemble the filtrate. Also, the more rapidly filtration occurs, the less time there is for the urine to be modified during its passage through the tubules.

Reabsorption from the proximal tubule. Reabsorption affects all the glucose of the filtrate, up to 70 percent of its water and sodium (the remainder is absorbed in the distal tubule), most of the potassium and chloride ions,

By courtesy of Andrew P Evan and Vincent H Gattone II

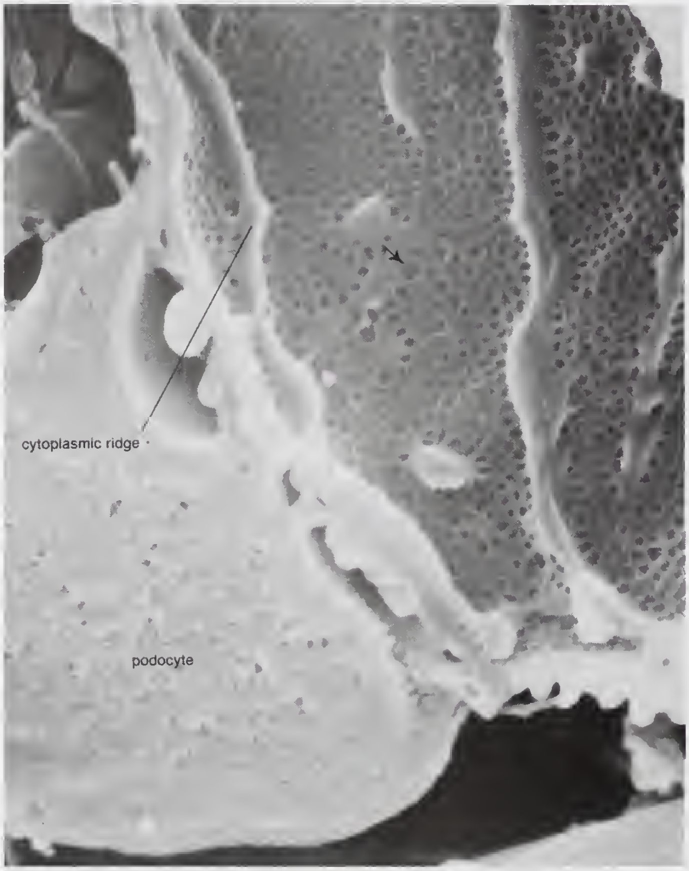


Figure 8: The inner surface of the glomerular endothelium with an associated podocyte. The endothelial cells are characterized by numerous large (500–1000 Å) fenestrae (arrow).

Table 2: Effect of Tubular Reabsorption on Urine
(illustrative 24-hour figures)

	glomerular filtrate	urine	tubular reabsorption (percent)
Water	170 l	1.5	99.1
Glucose	170 g	—	100
Sodium	560 g	5 g	99.1
Chlonde	620 g	9 g	98.5
Phosphate	5.1 g	1.2 g	76.5
Calcium	17 g	0.2 g	98.8
Urea	51 g	30 g	41.4
Sulfate	3.4 g	2.7 g	20.6

some of the uric acid, 40 percent of the urea, and little or none of the sulfate. Of the total solids 75 percent are reabsorbed in the proximal tubule. The first part of the tubule absorbs amino acids, glucose, lactate, and phosphate; the whole convolution absorbs sodium, potassium, calcium, and chloride and, by removing bicarbonate, acidifies the fluid slightly.

The tubule has only a certain capacity for reabsorption. Thus, normally all the glucose arriving in the filtrate is absorbed; but if plasma glucose is increased to high enough levels, the glucose arrives at the tubule cells faster than it can be absorbed—a condition that occurs in diabetes. In other words, there is a critical rate of delivery determined by plasma concentration and filtration rate, and a maximum reabsorptive capacity for each substance in the filtrate. The rate of tubular reabsorption has an upper maximum value that is constant for any given substance. Consequently, if the plasma level rises sufficiently, all surplus of the substance will pass out in the urine; this is true even for glucose, which is totally reabsorbed under normal conditions. On the other hand, the upper maximum value is much lower for phosphate, so there is normally always some phosphate in thc urine. The proximal tubular reabsorption of phosphate is also affected by the phosphate content of the filtrate and is influenced by parathyroid hormone. Phosphate competes with glucose for reabsorption, and its reabsorption is reduced by parathyroid hormone and by vitamin D and is increased, at least for some time, by a high dietary phosphate intake. The amino acids also have their own maximum tubular reabsorption values, but these are high enough to ensure that they are entirely reabsorbed under normal conditions; in certain rare inherited disorders such as cystinuria, in which there is excessive excretion of cystine, their rcabsorption is reduced.

Reabsorption of sodium

The reabsorption of about 70 percent of the sodium ions in the filtrate means that a similar value of water in the filtrate must accompany these ions as a vehicle to prevent a rising osmotic gradient (*i.e.*, to prevent a rising difference in the concentration of the sodium solution inside and outside the tubule). The energy required for the reabsorption of sodium into the blood uses 80 percent of the oxygen consumed by the kidney and represents one-eighth of the oxygen consumption of a person at rest. There is no evidence for active water transport, and the large volume of water reabsorption occurs passively in response to the movement of sodium. Since sodium is quantitatively the major osmotically active solute, the overall effect is to keep the fluid that remains in the tubular lumen, though much reduced in volume, roughly isosmotic with the original glomerular filtrate.

The active reabsorption of sodium (a positively charged ion) into the blood leaves the fluid remaining in the proximal tubule electronegative with respect to the peritubular fluids. This provides a driving force for the reabsorptive transport of negatively charged ions such as chloride, bicarbonate, and organic solutes. Reabsorption of neutral molecules such as urea into the blood is also driven by active sodium transport. Because the tubular epithelium is less permeable to urea and creatinine than it is to water or chloride, however, the free passive movement of water out of the tubular lumen leads to a rising luminal concentration of urea (*i.e.*, above the concentration in the original filtrate with plasma). As a result, a smaller proportion of filtered urea or creatinine than of sodium or water is reabsorbed into the blood, resulting in the elimination of a considerable amount in the urine.

Reabsorption from the loop of Henle. About one-third of the volume of the glomerular filtrate enters the descending limb of the loop of Henle. This fluid is isosmotic with plasma. The reabsorptive characteristics of the descending thin limb and those of the bend of the loop differ greatly from those of the ascending thick limb. The thin epithelium lining the thin limb is permeable to water and solute and has no power of active transport. Accordingly, the fluid entering the limb and the bend of the loop acquires the concentration of the fluid of the surrounding interstitial peritubular fluid. In contrast, the thick ascending limb lined by taller cells has low permeability to water and to urea but actively transports sodium and chloride into the peritubular fluid around both limbs. As a result this fluid in the medullary and deep cortical regions of the kidney becomes highly concentrated, reaching concentrations of up to four times that of the plasma (1,200 mosmoles per litre), mainly owing to the accumulation of sodium and chloride. This accumulation of solute, essential to the formation of a concentrated urine, is discussed in further detail below.

Reabsorption from the distal convoluted tubule. The active transport of sodium out of the ascending limb renders the fluid entering the distal convoluted tubule less concentrated than plasma. Active sodium reabsorption continues throughout the whole of the distal tubule, and this extends to thc early part of the collecting duct. As this part of the nephron is relatively impermeable to water, a large concentration gradient of sodium and chloride between the luminal fluid and the plasma is maintained, the concentration of sodium in the tubule being kept well below that of the plasma. The luminal fluid hcre is also markedly electronegative to the surrounding tissues. The mechanism of sodium reabsorption appears to bc directly linked to the secretion of potassium and of hydrogen ions into the tubule from the blood and is grcatly influenced by the hormone aldosterone, which is secreted by the adrenal gland when the body's sodium level is deficient.

The concentration of urine. As alrcady indicatcd, thc loop of Henle is critical to the ability of the kidney to concentrate urine. The high concentration of salt in the mcdullary fluid is believed to bc achicved in the loop by a process known as countercurrent exchangc multiplication. The principle of this process is analogous to the physical principle applicd in thc conduction of hot exhaust gases past cold incoming gas so as to warm it and conscrve heat. That exchange is a passive one; but in the kidney the countercurrent multiplier systcm uses energy to "pump" sodium out of the ascending limb of thc loop into the medullary fluid. From there it entcrs (by diffusion) the filtrate (isotonic with plasma) that is entering the descending limb from thc proximal tubule, thus raising its concentration a little abovc that of plasma. As this luminal fluid in turn reaches the ascending limb, and subsequently the distal tubule, it in turn provides more sodium to be pumped out into the surrounding fluid or blood, if necessary, and transported (by diffusion) back into the descending limb; this concentrating process continues until the osmotic pressure of the fluid is sufficient to balance the resorptive power of the collecting ducts in the medulla, through which all of the final urine must pass. This resorptive capacity in the ducts is regulated by antidiuretic hormone (ADH), which is secreted by the hypothalamus and stored in the posterior pituitary gland at the base of the brain. In the presence of ADH the medullary collecting ducts become freely permeable to solute and water. As a consequence the fluid entering the ducts (en route to the renal pelvis and subsequent elimination) acquires the concentration of the interstitial fluid of the medulla; *i.e.*, the urine becomes concentrated. On the other hand in the absence of ADH the collecting ducts are impermeable to solute and water; thus, the fluid in the lumen, from which some solute has been removed, remains less concentrated than plasma; *i.e.*, the urine is dilute.

Antidiuretic hormone

The secretion of ADH by the hypothalamus and its release from the posterior pituitary is part of a feedback mechanism responsive to the tonicity of plasma. This interrelation between plasma osmotic pressure and ADH output is mediated by specific and sensitive receptors at the base of the brain. These receptors are particularly sensitive to sodium and chloride ions. At normal blood tonicity there is a steady receptor discharge and a steady secretion of ADH. If the plasma becomes hypertonic (*i.e.,* has a greater osmotic pressure than normal), either from the ingestion of crystalloids such as common salt, or from shortage of water, receptor discharge increases, triggering increased ADH output, and more water leaves the collecting ducts to be absorbed into the blood. If the osmotic pressure of plasma becomes low, the reverse is the case. Thus water ingestion dilutes body fluids and reduces or stops ADH secretion; the urine becomes hypotonic, and the extra water is excreted in the urine.

The situation is complex because there are also receptors sensitive to changes in blood volume that reflexively inhibit ADH output if there is any tendency to excessive blood volume. Exercise increases ADH output and reduces urinary flow. The same result may follow emotional disturbance, fainting, pain, and injury, or the use of certain drugs such as morphine or nicotine. Diuresis is an increased flow of urine produced as the result of increased fluid intake, absence of hormonal activity, or the taking of certain drugs that reduce sodium and water reabsorption from the tubules. If ADH secretion is inhibited by the drinking of excess water, or by disease or the presence of a tumour affecting the base of the brain, water diuresis results; and the rate of urine formation will approach the rate of 16 millilitres per minute filtered at the glomeruli. In certain disorders of the pituitary in which ADH secretion is diminished or absent—*e.g.,* diabetes insipidus—there may be a fixed and irreversible output of a large quantity of dilute urine.

Tubular secretion. The only difference between secretory and reabsorptive tubular mechanisms lies in the direction of transport; secretory mechanisms involve the addition of substances to the filtrate from the plasma in the peritubular capillaries. The small amount of secretion that does occur, except for the secretion of potassium and uric acid, takes place in the proximal tubule. Hydrogen ions are also secreted and ammonia is generated, but they are special cases and are discussed below under *Regulation of acid-base balance.* As in the case of reabsorption, secretion occurs both passively and actively against an electrochemical gradient.

Several drugs are actively secreted, and some of these appear to share a common pathway so that they may compete with each other for a limited amount of energy. This may be turned to therapeutic advantage in the case of penicillin, which is eliminated partly by tubular secretion. The drug probenecid, which can be given simultaneously, competes with penicillin at its secretory site and thus helps to raise the level of penicillin in the blood in the treatment of certain infections. Endogenous (originating within the body) compounds that are secreted also include prostaglandins, bile salts, and hippurate. Uric acid derived from nucleoproteins freely passes the glomerular barrier and is normally largely reabsorbed in the proximal tubule. In some circumstances, however, it is also secreted by other parts of the same convoluted tubule.

The secretion of potassium by the distal tubule is one of the most important events in the kidney as its control is fundamental to the maintenance of overall potassium balance. More than 75 percent of the filtered potassium is reabsorbed in the proximal tubule and in the ascending limb of the loop of Henle, and this percentage remains virtually constant, irrespective of how much is filtered. The amount eliminated in the urine, which is ultimately determined by the dietary intake, is controlled by the distal convoluted tubule. In persons consuming a normal diet, probably about 50 percent of the urinary potassium is secreted into the urine by the distal tubules; this amount can be adjusted according to body need. One of the several factors that influence potassium secretion is a hormone secreted by the cortex of the adrenal gland, aldosterone. In the absence of aldosterone and other mineralocorticoids (adrenocortical steroids affecting electrolyte and fluid balance), potassium secretion is impaired, and potentially dangerous amounts can accumulate in the blood. Excess aldosterone promotes potassium excretion.

Regulation of acid-base balance. The cells of the body derive energy from oxidative processes that produce acidic waste products. Acids are substances that ionize to yield free protons, or hydrogen ions. Those hydrogen ions that derive from nonvolatile acids—such as lactic, pyruvic, sulfuric, and phosphoric acids—are eliminated in the urine. The kidney contains transport mechanisms that are capable of raising the concentration of hydrogen ions in the urine to 2,500 times that in the plasma or, when appropriate, lowering it to one-quarter that of the plasma.

Theoretically, acidification of urine could be brought about either by the secretion of hydrogen ions into the tubular fluid or by the selective absorption of a buffer base (a substance capable of accepting hydrogen ions; *e.g.,* filtered bicarbonate). Current evidence indicates that both filtration and secretion are essential to hydrogen ion excretion and that both proximal and distal convoluted tubules are involved.

The bi-carbonate threshold

The bulk of the bicarbonate filtered at the glomerulus is reabsorbed in the proximal tubule, from which it passes back into the peritubular capillaries. This mechanism is designed to keep the normal plasma bicarbonate concentration constant at about 25 millimoles per litre. When the plasma concentration falls below this level, no bicarbonate is excreted and all filtered bicarbonate is reabsorbed into the blood. This level is often referred to as the bicarbonate threshold. When the plasma bicarbonate rises above 27 millimoles per litre, bicarbonate appears in the urine in increasing amounts.

The brush borders of the cells of the proximal tubules are rich in the enzyme carbonic anhydrase. This enzyme facilitates the formation of carbonic acid (H_2CO_3) from CO_2 and H_2O, which then ionizes to hydrogen ions (H^+) and bicarbonate ions (HCO_3^-). The starting point for bicarbonate reabsorption is probably the active secretion of hydrogen ions into the tubular fluid. These ions may be formed under the influence of carbonic anhydrase from CO_2 liberated from oxidation of cell nutrients and H_2O already in the cells. The filtered base, bicarbonate, accepts the hydrogen ions to form carbonic acid, which is unstable and dissociates to form CO_2 and H_2O. The partial pressure of CO_2 in the filtrate rises, and, as CO_2 is highly diffusible, it passes readily from the tubular fluid into the tubular cells and the blood, and the water is either dealt with in the same way or is excreted. In the meantime the proximal tubular cells are actively reabsorbing filtered sodium, which is balanced by the HCO_3^- formed within the cells from the CO_2 generated by the hydrogen ions in the luminal fluid. Thus the bicarbonate actually reabsorbed is not that which was originally the filtrate, but the net effect is the same as if this were the case.

Other bases besides HCO_3^- may buffer the hydrogen ions secreted into the distal tubules; in addition, the ions may combine with ammonia also secreted by the tubules. The most important non-bicarbonate base present in the filtrate is dibasic phosphate (Na_2HPO_4), which accepts hydrogen ions to form monobasic phosphate (NaH_2PO_4). A measure of the amount of hydrogen ion in the urine that is buffered by bases such as bicarbonate and phosphate is made by the titration of urine with strong base until the pH of the plasma from which the filtrate is derived (7.4) is achieved. This is called the titratable acidity of urine and usually amounts to between 20 and 40 millimoles of H^+ per day.

In normal circumstances about two-thirds of the hydrogen ions to be secreted in the urine is in the form of ammonium salts (*e.g.,* ammonium chloride). Ammonia (NH_3) is not present in plasma or filtrate but is generated in the distal tubular cells and passes into the lumen probably by passive diffusion down a concentration gradient. In the lumen the NH_3 combines with hydrogen ions secreted into the tubule to form ammonium ions (NH_4^+), which are then trapped in the lumen because the lipid walls of the tubular cells are much less permeable to the charged than to the uncharged molecules.

It is now known that ammonia is formed from the hydrolysis of glutamine (an amino acid) to form glutamic acid and ammonia by the enzyme glutaminase. A further molecule of ammonia is obtained by the deamination of glutamic acid to form glutaric acid, which is then metabolized. The more acidic the urine is, the greater is its content of ammonium ions; the introduction of hydrogen ions (*e.g.*, from the diet) stimulates production of ammonium by the tubular cells. The ammonium is excreted in the urine as ammonium salts of surplus anions (negative ions) such as chloride, sulfate, and phosphate, thus sparing for retention other cations (positive ions) such as sodium or potassium.

In summary, hydrogen ion secretion can be considered in three phases. The first occurs in the proximal tubule, where the net result is tubular reabsorption of filtered bicarbonate. The second and third phases take place in the distal tubule, where monobasic phosphate and ammonium salts are formed. The total tubular cell secretion of hydrogen ion is therefore the sum of titratable acidity, the amount of ammonium ion excreted, and the amount of bicarbonate ion reabsorbed. The last may be assessed by calculating the amount of bicarbonate filtered (*i.e.*, plasma concentration of bicarbonate × glomerular filtration rate and subtracting any bicarbonate excreted in the urine). Total hydrogen ion secretion normally amounts to 50–100 millimoles per day but may rise considerably above this in disorders associated with excess acid production, such as diabetes.

Volume and composition. The volume and composition of normal urine vary widely from day to day, even in healthy individuals, as a result of food and fluid intake and of fluid loss through other channels as affected by environmental conditions and exercise. The daily volume averages 1.5 litres (about 1.6 quarts) with a range of 1–2.5 litres, but after copious sweating it may fall as low as 500 millilitres, and after excess fluid intake it may reach three litres or more. There is also variation within a 24-hour period. Excretion is reduced in the early hours, maximal during the first few hours after rising, with peaks after meals and during the early stages of exertion. The urine produced between morning and evening is two to four times the night volume. The excessive secretion of urine (polyuria) of chronic renal disease is typically nocturnal.

The volume of urine is regulated to keep plasma osmotic concentration constant, to control the total water content of the tissues, and to provide a vehicle for the daily excretion to the exterior of some 50 grams of solids, mostly urea and sodium chloride. In a man who ingests 100 grams of protein and 10 grams of salt daily, the urine will contain 30 grams of urea and 10 grams of salt; there are many other possible constituents, but they amount to less than 10 grams overall.

Some urinary constituents (Table 3)—the products of metabolism of nitrogenous substances obtained from food—vary widely in relation to the composition of the diet; thus the excretion of urea and sulfate is dependent on the diet-protein content. A high-protein diet may yield a 24-hour output of 17 grams of nitrogen, a low-protein diet of the same calorific value only three to four grams.

The urine is normally clear. It may be turbid from calcium phosphate, which clears if acetic acid is added. Microscopic deposits include occasional casts, vaguely resembling in form the renal tubules from whose lining they have been shed. An ammoniacal smell is the result of decomposition of urea to ammonia by bacteria and is commonly present on babies' diapers. Certain foods and drugs may cause distinctive odours. The colour of urine depends on its concentration but is normally a bright clear yellow from the pigment urochrome, an end product of protein metabolism. There are also traces of other pigments: urobilin and uroerythrin. The colour may be influenced as well by food dyes, beetroot, and certain drugs.

Table 3: Some Urine Constituents (g/24 hours)

Urea	25–30
Uric acid	0.6–0.7
Creatinine	1.0–1.2
Hippuric acid	0.7
Ammonia	0.7
Amino acids	3.0
Sodium	1–5 (NaCl 15.0)
Potassium	2–4
Calcium	0.2–0.3
Magnesium	0.1
Chloride	7
Phosphate	1.7–2.5
Sulfate	1.8–2.5

The specific gravity of urine may vary between 1.001 and 1.04 but is usually 1.01–1.025. Such variation is normal, and a fixed low specific gravity is an indication of chronic renal disease. If fluid intake is stopped for 24 hours, even a normal kidney will secrete urine with a specific gravity of at least 1.025. There is a limit to the concentrating powers of the kidney, so that the urine is rarely more than four times as concentrated as plasma. In order to excrete their normal solute load, the kidneys need a minimum water output of 850 millilitres as a vehicle; this volume is often called the minimum obligatory volume of urine. If this is not available from intake it has to be withdrawn from the tissues, causing dehydration; but the usual intake is well above the minimum and the urine is rarely at its maximum possible concentration. The reaction of the urine is usually acid, with an overall range of pH 4 to 8 (lemon pie has a pH of 2.3; the value 8 is slightly alkaline, about equal to the pH of a 1 percent solution of sodium bicarbonate).

Abnormal constituents of urine

Foreign proteins of molecular weight less than 68,000 are excreted in the urine, while those of the plasma are retained in the body. If, however, the kidneys are damaged by disease or toxins, the glomeruli will transmit some of the normal serum albumin and globulin and the urine will coagulate on warming. Normally, the urine contains only very small amounts of protein (less than 50 milligrams per 24 hours); however, protein content in the urine is increased after exercise, in pregnancy, and in some persons when standing (orthostatic albuminuria). The protein loss may be greatly increased in certain chronic renal diseases; in the nephrotic syndrome it may even reach 50 grams in a 24-hour period. Certain specific and easily identifiable proteins appear in the urine in diseases associated with the overgrowth of cells that make immunoglobulins.

Glucose is found in the urine in diabetes mellitus. In some healthy persons, however, there may also be an abnormal amount of glucose in the urine because of a low threshold for tubular reabsorption, without any disturbance of glucose metabolism. Lactosuria (abnormal amount of lactose in the urine) may occur in nursing mothers. Ketone bodies (acetone, acetoacetic acid) are present in traces in normal urine but in quantity in severe untreated diabetes and in relative or actual carbohydrate starvation; *e.g.*, in a person on a high-fat diet.

The urine may contain hemoglobin or its derivatives after hemolysis (liberation of hemoglobin from red blood cells), after incompatible blood transfusion, and in malignant malaria (blackwater fever). Fresh blood may derive from bleeding in the urinary tract. Bile salts and pigments are increased in jaundice, particularly the obstructive variety; urobilin is greatly increased in certain diseases such as cirrhosis of the liver.

Porphyrins are normally present only in minute amounts but may be increased in congenital porphyria, a disease characterized by sensitivity to sunlight or by insanity. The presence of porphyrins also may increase after ingestion of sulfonamides and some other drugs.

The normally small quantities of amino acids in the urine may be much increased in advanced liver disease, in failure of tubular reabsorption, and in certain diseases due to inborn errors of protein metabolism. Phenylketonuria, a disease identified by the presence of phenylpyruvic acid in the urine, is due to lack of the enzyme phenylalanine hydroxylase, so that phenylalanine is converted not to tyrosine but to phenylpyruvic acid. The presence of this acid in blood and tissues causes mental retardation; if the urine of every newborn infant is tested, restriction of phenylalanine in the diet in such cases may be beneficial. Alkaptonuria, a disease identified by the presence of homogentisic acid in the urine, is due to lack of the enzyme that catalyzes the oxidation of homogentisic acid; deposits of the acid in the tissues may cause chronic arthritis or spinal disease. Other such disorders are cystinuria, the presence of the amino acid cystine in the urine, when the bladder may contain cystine stones; and maple syrup disease, another disorder involving abnormal levels of amino acid in the urine and blood plasma.

URINE COLLECTION AND EMISSION

Passage of urine from nephron to bladder

From the nephrons the urine enters the final 15 or 20 collecting tubules that open on to each papilla of the renal medulla, projecting into a minor calyx. These open into two or three major calyxes, and these in turn open into the renal pelvis, the upper expanded portion of the ureter.

Urine is passed down the channel of the renal pelvis and ureter by a succession of peristaltic waves of contraction that begin in the muscle fibres of the minor calyxes, travel out to the major calyxes and then along the ureter every 10–15 seconds. Each wave sends urine through the ureteric orifice into the bladder in discontinuous spurts; these can be seen through a cystoscope if a dye is injected into the bloodstream. Gravity aids this downward flow, which is faster when one is standing erect. Though the overall picture suggests that there is a pacemaker (a set of specialized cells capable of rhythmic contractions) near the pelviureteric junction, this has never been satisfactorily demonstrated in the tissue. The pressure in the renal pelvis is normally low, but the smooth muscle coat of the ureter is a powerful one and the pressure above an obstructed ureter may rise as high as 50 millimetres of mercury. The ureters are doubly innervated from the splanchnic nerves above and the hypogastric network below.

The bladder. The bladder is a hollow organ of variable capacity, with a powerful intermediate muscle coat that empties the organ when it contracts, and two muscular sphincters that keep the exit closed at all other times. This smooth muscle coat constitutes the powerful detrusor muscle. At the base of the bladder the region of the bladder neck, or trigone, is demarcated by the two ureteric orifices and the internal opening of the urethra. Muscle fibres loop around the urethral opening to form the internal sphincter, which is under involuntary control. The external sphincter consists of two striated muscles under voluntary control: the compressor urethrae, which surrounds the membranous urethra, and the pair of bulbocavernosus muscles.

The mucous membrane lining the bladder is distensible; it is ridged in the empty organ and smoothed out in distension. In micturition the longitudinal muscle of the bladder shortens to widen the bladder neck and allow urine to enter the urethra. The urethra normally contains no urine except during the act of micturition, its walls remaining apposed by muscle tone. In the male, but not in the female, the external sphincter can maintain continence even if the internal sphincter is not functioning.

Innervation of bladder and urethra

The innervation of the bladder and urethra is complex and important. Essentially, there are three groups of nerves: (1) The parasympathetic nerves constitute the main motor supply to the detrusor; they make it contract, raise pressure within the bladder, relax the internal sphincter, and cause emptying. Afferent parasympathetic channels convey impulses from stretch receptors in the bladder wall to higher centres, permitting cognizance of the state of distension of the organ and stimulating the desire to micturate. (2) The sympathetic nerves stimulate closure of the ureteric and internal urethral orifices and contraction of the internal sphincter, and their action on the detrusor is inhibitory; *i.e.,* the effect is to prevent bladder outflow. Thus the sympathetic serves to control the situation in the distending bladder up to the point when evacuation can be deferred no longer. Afferent paths in the sympathetic convey sensations of pain, overdistension, and temperature from the mucosa of the bladder and the urethra. (3) The somatic nerves cause contraction of the external sphincter; their sensory fibres relay information as to the state of distension of the posterior urethra.

Both the parasympathetic nerves and the somatic nerves (pudendal nerve) to the external sphincter relay impulses to the second through fourth sacral segments of the spinal cord, which constitute a reflex centre for the control of bladder function. This centre connects with higher centres in the brain by ascending and descending fibres in the spinal cord.

Bladder function in micturition. Certain reflexes combine to ensure both maintenance of a steady holding state for urine and normal progressive micturition with complete emptying. When the internal pressure of the bladder rises, it contracts; and it also contracts when urine enters the urethra.

Both bladder sphincters are normally closed. As the organ fills with urine, the contractile response of the muscle wall causes a rise in internal pressure. Relaxation then occurs as an active process of adjustment so that the organ may hold its contents at a lower pressure. As urine continues to enter the bladder, this rise and fall of pressure continues in steplike fashion, with the final pressure always gradually rising.

The repeated transient contraction waves at first are small and are not consciously felt; later, stimuli reach the brain and cause pain and a sharp rise of pressure. These later major contractions can be inhibited voluntarily. The desire to micturate begins at around a content of 400 millilitres, but it can be voluntarily overridden until the content reaches 600–800 millilitres, equivalent to a pressure of 100 millimetres of water. Until this point the sphincters remain contracted to keep the urethral exit closed, but eventually the desire to micturate becomes urgent and irrepressible. Until that time, voluntary inhibition of the detrusor and contraction of the perineal muscles have kept the internal pressure as low as possible and have prevented efflux. The threshold is dependent to some extent on the rate of filling and is higher when filling is slow; and training affects the amount the bladder can retain. In young children the situation is less controllable, and even small amounts of urine may excite reflex evacuation. Emotional influences are important. Anxiety inhibits the capacity of the bladder to relax on filling, so that under conditions of stress there may be some involuntary passage of small quantities of urine.

Normal micturition

Micturition. Micturition is a complex activity, partly reflex and unconscious and mediated by the lower spinal cord centres, and partly under conscious control by the higher centres of the brain. Voluntary micturition begins with willed messages from the brain that reach the bladder via the motor fibres of the pelvic nerves to stimulate the detrusor, at the same time actively relaxing both urethral sphincters. But the reflexes already mentioned ensure that, once the process has begun and urine has entered the urethra, the contraction of the detrusor will continue and the sphincters will remain relaxed until evacuation is complete and the bladder empty. Evacuation is aided by voluntary contraction of a wide range of accessory muscles. The muscles of the abdominal wall contract to increase pressure on the bladder from without; the diaphragm descends and the breath is held; at the same time there is relaxation of the muscles of the perineal floor. Thus voluntary initiation and control of micturition is effected partly by an active process of stimulating parasympathetic sacral nerve outflow, partly by removing the normal inhibition exerted by the higher centres on the reflex centres in the spinal cord. Once begun, micturition is carried through to completion by lower and higher centres acting in concert; sensory messages from the urine-distended urethra also play a part. It follows that even if a bladder is not particularly distended and if reflex emptying is not urgent, the bladder can nevertheless be evacuated by voluntary contraction of the abdominal wall, so initiating the reflex process that, once begun, takes over.

TESTS OF RENAL FUNCTION

Quantitative tests. Important quantitative tests of renal function include those of glomerular filtration rate, renal clearance, and renal blood flow. Tests are also made to estimate maximal tubular activity, tubular mass, and tubular function. Radiological and other imaging methods are useful noninvasive diagnostic techniques, and renal biopsy is valuable in detecting pathological changes that affect the kidneys. In both clinical and experimental studies one of the most fundamental measures of renal function is that of the glomerular filtration rate (GFR). The GFR is calculated by measuring the specific clearance from the body of a substance believed to be excreted solely by glomerular filtration. The renal clearance of any substance is the volume of plasma containing that amount of the substance that is removed by the kidney in unit time (*e.g.*, in one minute). Clearance, or the volume of plasma cleared, is an artificial concept since no portion of the plasma is ever really cleared in this fashion.

Renal clearance

It was soon realized, however, that if a substance could be found that was freely filtered by the glomeruli and was neither reabsorbed, metabolized, nor secreted by the renal tubules, its clearance would equal the GFR. This is so in these circumstances because the amount of such a substance excreted in the urine in one minute would equal the amount that has been filtered at the glomeruli in the same time. If the concentration of the substance in the plasma (which is the same as that in the glomerular filtrate) is known, the clearance volume must represent the volume of glomerular filtrate.

The first substance identified to be excreted in this way was the polysaccharide inulin (molecular weight about 5,000), which is extracted from the roots of dahlias. Although inulin is not naturally found in human plasma it is nontoxic and can be injected or infused into the bloodstream. Its concentration also can be measured readily and accurately. In the adult male the GFR is 125 millilitres per minute per 1.73 square metres of body surface. In the adult female, the values are about 85 percent of those for the same standard area of body surface. Inulin clearance is now accepted as the standard for estimation of the GFR.

Clearance value is not the same as excretion rate. The clearance of inulin and some other compounds is not altered by raising its plasma concentration, because the amount of urine completely cleared of the agent remains the same. But the excretion rate equals total quantity excreted per millilitre of filtrate per minute, and this value is directly proportional to its plasma concentration.

Substances, such as urea, whose clearance is less than the GFR must be reabsorbed by the renal tubules, while substances whose clearance is greater than the GFR must be secreted by the renal tubules. Since the discovery of inulin, researchers have identified a small number of other substances that are excreted by the kidney in a similar fashion and that have similar clearance values. These include vitamin B_{12}, circulating free in plasma and unbound to protein, and sodium ferrocyanide.

The clearance of creatinine was used as a measure of renal function before inulin was discovered; because this substance is found naturally in plasma, creatinine clearance is still widely used as an approximate measure of the GFR. Creatinine is produced in the body at virtually a constant rate, and its concentration in the blood changes little; accordingly, creatinine clearance is usually measured over a period of 24 hours. There is evidence that in humans creatinine is secreted into the urine by renal tubules as well; however, the amount is small and constant and has little effect on the measure of the GFR.

The concept of clearance is also useful in the measurement of renal blood flow. Para-aminohippuric acid (PAH), when introduced into the bloodstream and kept at relatively low plasma concentrations, is rapidly excreted into the urine by both glomerular filtration and tubular secretion. Sampling of blood from the renal vein reveals that 90 percent of PAH is removed by a single circulation of blood through the kidneys. This high degree of PAH extraction by the kidney at a single circulation implies that the clearance of PAH is approximately the same as renal plasma flow (RPF). The 10 percent of PAH that remains in renal venous blood is conveyed in blood that perfuses either nonsecretory tissue, such as fibrous tissue or fat, or parts of the tubule that do not themselves secrete PAH. In practice this small remaining percentage is usually ignored, and the clearance of PAH is referred to as the effective renal plasma flow. In humans PAH clearance is about 600 millilitres per minute, and thus true renal plasma flow is about 700 millilitres per minute.

Estimation of the GFR and RPF allows the proportion of available plasma perfusing the kidney that is filtered by the glomerulus to be calculated. This is called the filtration fraction and on average in healthy individuals is 125/600, or about 20 percent. Thus about one-fifth of plasma entering the glomeruli leaves as filtrate, the remaining four-fifths continuing into the efferent glomerular arterioles. This fraction changes in a number of clinical disorders, notably hypertension.

Filtration fraction

Reference has already been made to the fact that the renal tubules possess a limited capacity to perform certain of their functions. This is the case, for example, in their ability to concentrate and dilute urine and to achieve a gradient of hydrogen ions between urine and blood. Concentrating power can be tested by depriving the individual of water for up to 24 hours, or, more simply, by introducing a synthetic analogue of ADH into each nostril. The water deprivation test assesses the individual's capacity to produce ADH and the sensitivity of the renal concentrating mechanism to circulating ADH. The use of an analogue of ADH assesses only the sensitivity of the renal tubules to the hormone.

The limits of renal ability to excrete acid and establish a gradient of the concentration of hydrogen ions between plasma and urine has been mentioned above. The power of acidification of urine is best estimated by measuring the pH of urine after the administration of ammonium chloride in divided doses over two or three days. Other specific functions that are tested include the individual's ability to conserve sodium, potassium, and magnesium. In general, these tests are carried out by administering diets that are deficient in these electrolytes and then estimating the minimum rate of excretion after several days.

Radiological and other imaging investigations. Imaging techniques are used to determine the anatomical site, configuration, and level of functioning of the kidneys, pelvis, and ureters. A plain X ray nearly always precedes any other more elaborate investigation, so that the size, outline, and position of the two kidneys, as well as information about the presence or absence of calcium-containing renal stones or zones of calcification can be ascertained. Excretion urography is one of the simplest methods of defining these aspects more precisely, though this radiological method is giving way to noninvasive imaging methods such as ultrasonography and nuclear magnetic resonance (NMR). In excretion urography, the kidneys are observed in X rays after intravenous injection of a radiopaque iodine-containing compound that is excreted largely by glomerular filtration within one hour of the injection. A series of X-ray images (nephrograms) then indicates when the contrast substance first appears and reveals the increasing radiographic density of the renal tissue. The X rays also indicate the position, size, and presence of scarring or tumours in the organs and provide an approximate comparison of function in the two kidneys. Finally the dye collects in the bladder, revealing any rupture or tumour in this organ.

Excretion urography

Obstruction to the flow of urine also may be revealed by distension of the calyceal system above the site of obstruction. This is more clearly detected by urography, in which contrast medium is injected through a fine catheter introduced either directly into the pelvis of the kidney or into the ureteral orifice visualized during cystoscopy. A micturating cystogram involves the injection of contrast substance into the bladder and is of importance in the investigation of urinary tract infection in childhood. It may show the reflux of urine from the bladder upward into the ureters or kidneys on micturition. Because of the risk of radiation to the gonads this test should be conducted only on certain patients.

A radioactive renogram involves the injection of radioactive compounds that are concentrated and excreted by the kidney. The radiation can be detected by placing gamma scintillation counters externally over the kidneys at the back; the counts, transcribed on moving graph paper, yield characteristic time curves for normal and disordered function.

A picture of renal circulation can be obtained by introducing a radiopaque substance directly into the abdominal aorta just above the origin of the renal arteries, or directly into the renal arteries themselves. The contrast material yields a renal angiogram, showing the renal vascular tree. The technique is especially valuable in demonstrating the presence of localized narrowing or obstructions in the circulation or of localized dilatations (aneurysms). Tumours, which tend to be well vascularized, are also distinguishable from cysts, which are not well supplied with blood.

Non-invasive imaging techniques

Ultrasound and NMR have the advantage of being noninvasive and apparently free from risk to the patient. They are useful in detecting tumours of the kidney or adjacent structures and in distinguishing tumours from cysts. Ultrasound techniques are comparatively simple and have replaced other methods in detecting the presence of polycystic kidneys.

Renal biopsy. The visual, usually microscopic, examination of a specimen of kidney tissue removed from a living patient (renal biopsy) is the only investigative method that yields exact histological data on renal structure. The material for examination is usually obtained by inserting a special needle through the skin of the back into the kidney substance and withdrawing a fragment of tissue. A general anesthetic is not usually required, the procedure occupying only a few minutes. Renal biopsy has been valuable in clarifying several renal disorders, notably those affecting the glomeruli, and in revealing their prognosis and natural course. The only serious potential complication is excessive bleeding, but this is rare. The procedure is not justified, however, if the patient possesses only one kidney or suffers from a bleeding disorder or from severe, uncontrolled high blood pressure.

THE ROLE OF HORMONES IN RENAL FUNCTION

Certain hormones and hormonelike substances are intimately related to renal function. Some of these, such as ADH (or vasopressin), are produced outside the kidney and travel to the kidney via the blood as chemical messengers. Others are produced within the kidney and appear to exert only a local effect. The role of ADH in controlling diuresis has already been discussed. ADH regulates water excretion by increasing the permeability of the collecting ducts to water and salt and by accelerating water and ion transfer in a direction determined by the osmotic gradient. The receptors at the base of the brain form part of the feedback mechanism that (1) stimulates ADH output if the osmotic concentration of extracellular fluid (ECF) is high, so as to concentrate the urine, and (2) reduces ADH output and so dilutes the urine if osmotic concentration of ECF and of plasma falls.

Hormones of the adrenal cortex

The hormones of the adrenal cortex are also important in influencing renal function, directly or indirectly. In stress situations, as after an injury or a surgical operation, the output of hydrocortisone and other corticosteroids is increased because the adrenals are stimulated by adrenocorticotropin (ACTH), a secretion of the pituitary gland. Hydrocortisone increases protein breakdown, and consequently the output of nitrogen in the urine, and affects water metabolism; lack of hydrocortisone reduces the power of the kidney to deal with normal water loads. The hormone also promotes sodium retention and loss of potassium and hydrogen ions by the kidney. Aldosterone influences electrolyte metabolism by facilitating the reabsorption of sodium ions at the distal tubules, also at the expense of hydrogen and potassium excretion. The action of aldosterone has been described as priming the sodium reabsorption pump; it is the adrenal hormone most important to tubular function. It also influences the ability of the bowel to absorb sodium, and thus its level of production profoundly influences overall sodium balance. Deficiency of aldosterone allows a steady loss of sodium in the urine, causing a fall in blood pressure that may result in fainting.

The action of the parathyroid glands is to increase blood calcium by mobilizing calcium from the bones and other sourccs; if this hormone functions to excess, as in tumours of the glands, the urinary loss of calcium is much increased and calcium stones tend to form in the kidneys and the bladder. Parathyroid hormone also increases the renal excretion of phosphate and accelerates the conversion of hydroxylated vitamin D to the dehydroxylated form in the kidney. The pituitary growth hormone facilitates protein synthesis and decreases the urinary loss of nitrogen. The sex hormones estrogen and progesterone exert an ill-defined activity as regards salt and water metabolism.

Renin–angiotensin system

The juxtaglomerular apparatus (JGA), consisting of an asymmetrical cuff of large granular cells in the wall of the afferent arteriole near its entry into the capsule of the nephron, contains renin in the granules in the cells. Renin is a true internal secretion of the kidney. Entering the plasma, it acts as an enzyme that induces one of the plasma globulins to yield angiotensin I, which is inactive, and which gives rise in turn to angiotensin II, the most potent agent for constricting the blood vessels and raising the blood pressure. The formation of renin at the JGA is induced by a fall in blood pressure and inhibited by a rise. When the pressure falls, the output of angiotensin II raises the pressure and also excites the release of aldosterone from the adrenal cortex. This process is another example of a feedback mechanism analogous to that controlling the output of ADH.

Prostaglandins

Among the prostaglandins, a group of hormonelike fatty acids synthesized throughout the body, the ones found in the kidney tissues appear to exert local influence on various aspects of renal function. Unlike true hormones, prostaglandins are not transported away from their site of origin by the blood. The interstitial and collecting duct cells of the kidney produce a characteristic prostaglandin, PGE_2, and the renal cortex produces PGI_2, or prostacyclin. Renal prostaglandins interact with the renin–angiotensin system in several ways. The renal cortex prostaglandin PGI_2 mediates the increased release of renin in response to decreases in renal blood flow. The angiotensin subsequently formed in the plasma stimulates production of the interstitial and duct cell prostaglandin (PGF_2), which itself inhibits angiotensin-induced vasoconstriction. For this reason the renal cortex prostaglandin is thought to be an important vasodilator, maintaining renal blood flow when this is threatened (for example, after blood loss). Prostaglandins may also inhibit the action of ADH on the distal tubule and collecting ducts, and the interstitial and duct cell prostaglandin may have a direct effect in inhibiting renal tubular sodium reabsorption; however, the relative importance of these different actions in the healthy human is not known.

Another substance that causes the dilation of blood vessels, the enzyme kallikrein, may also exert an influence on renal blood flow. Kallikrein is secreted by renal tubules and is added to the urine in the distal tubules. It activates the conversion of kininogen to bradykinin, which is also a powerful vasodilator. Bradykinin is inactivated by a kininase, which also converts angiotensin I to angiotensin II, a substance that causes the constriction of blood vessels. Thus the same enzyme that inactivates the vasodilator bradykinin catalyzes the production of the vasoconstrictor angiotensin II. This relationship again suggests a delicately balanced internal control system.

Dopamine is a putative renal hormone that may affect salt balance. The sympathetic nerves that travel to the kidney, the terminals of which release catecholamines such as norepinephrine, are not believed to be important in controlling tubular salt reabsorption. Transplanted human kidneys function adequately despite the lack of any nerve supply and so renal nerves are not essential. However, because dopamine (also a catecholamine released at sympathetic nerve endings) is present in urine in amounts far in excess of the amount that might be filtered from the blood, it may be deduced that some dopamine is formed within the kidney. It is now believed that dopamine is formed enzymatically within the kidney from its precursor, L-dopa, which freely circulates in the blood, and that only small amounts are released by sympathetic nerve endings. Dopamine is a powerful natriuretic substance (*i.e.*, one capable of increasing urinary salt loss) and renal vasodilator. Its role in salt balance, renal function, and blood pressure control remains speculative.

The most recently identified hormone that influences renal function is secreted by special "stretch receptor" cells in the atria of the heart in response to a rise in atrial pressure, as during heart failure. This hormone, called atrial natriuretic peptide (ANP), exerts a vasodilator effect on the kidney and also reduces tubular reabsorption of sodium. Both actions result in increased urinary elimination of salt and water and tend to restore atrial pressure toward the normal. It is probably an important hormone controlling the volume of the extracellular fluid.

BIOLOGICAL CONSIDERATIONS

During most of the pregnancy period the glomerular filtration rate (GFR) is increased by as much as 50 percent, corresponding to an increase in renal blood flow of up to 25 percent in the middle three months of pregnancy. Glycosuria is frequent and is due to increased glucose loading of the filtrate; there is some sodium retention with a tendency to abnormal accumulation of serous fluid (edema), and some protein may appear in the urine. Anatomical changes include enlargement and dilation of the pelvis and ureters, caused by both hormonal action and partial ureteric obstruction by the gravid uterus. These changes may be responsible for the increased susceptibility to urinary tract infection during pregnancy.

The kidneys of the fetus begin to function well before birth, as indicated by a steady rise in the urea and uric acid content of the amniotic fluid in which the fetus exists; the fetus probably swallows fluid and voids it as urine. But even at birth, half the work of excretion is still being carried out via the placental circulation and the maternal kidneys, and this dependence is abruptly curtailed. Kidney function is far from fully developed in the newborn infant. The glomerular filtration rate is only some 30 millilitres per minute per square metre of body surface, compared to 75 in the adult, and tubular function does not attain adult performance until the end of the first year. The 24-hour output of urine is only some 20 millilitres; the output of water and the renal clearance of sodium, potassium, and phosphate is low; the urine is dilute and often contains protein. Because the kidney has such a poor capacity to excrete solids, the infant is exposed to the dehydrating effect of vomiting and diarrhea, which readily induce renal failure.

Rhythms in urine output

There is an increased urine output at the commencement of muscular exercise, due to the general stimulation of circulation, but a later falling off with the fatigue and sweating caused by severe prolonged exertion. The 24-hour rhythm in output has been mentioned. The small output in the early morning hours is a practical convenience to prevent disturbance of sleep. If the natural sleep rhythm is inverted, as by working on night shift, electrolyte and water output follow suit. The urine is acidic at night and becomes less so, or alkaline, on rising. Output is maximal during the first waking hours and rises after meals. Because of all this variation in water and solute output, any analytic study of urine components must be conducted on specimens obtained over a 24-hour period.

(D.LeV./J.S.Ro.)

Excretory system diseases and disorders

The impact of diseases on the elimination of wastes and the conservation of an appropriate amount and quality of body fluid is the primary concern in this section. Many of the manifestations of renal disease can be accounted for in terms of disturbance of these two functions, and the alleviation of symptoms in those renal diseases that cannot yet be cured depends on knowledge of how these two functions are affected.

The eliminatory process does not, of course, end with the formation of urine; the urine has to pass down the ureters to the bladder, be stored there, and voided, usually under voluntary control. The whole mechanism can be deranged by structural changes in the lower urinary tract, by infection, or by neurological disorders that lead to abnormal emptying of the bladder. Disturbance of the lower urinary tract is an important cause of pain and distress, notably during pregnancy and in the elderly; and it can lead to serious and progressive damage to the kidneys, either by interfering with the drainage of urine or by allowing bacterial infection to have access to the kidney.

FUNCTIONAL ASPECTS

Effects of abnormal renal function on body fluid. Renal disease in its diverse forms can lead to bodily deficits or excesses of water, sodium, potassium, and magnesium, and also to protein deficits occasioned by great losses of protein in the urine. Inability of the kidney to function normally may lead to retention in the blood of the waste products of protein metabolism, such as urea and uric acid, and of other nitrogenous compounds such as creatinine. There may be abnormally high levels of phosphates in the blood, which in turn can lead (for reasons about which there is still some disagreement) to low blood levels of calcium. The calcium deficiency can cause tetany, a condition marked by muscular spasms and pain, and calcium may be lost from the bones in the process of restoring normal calcium levels in the blood and tissue fluid. For descriptive purposes, changes in volume, changes in composition, and protein depletion of renal origin will be discussed separately, but these disturbances can and often do coexist.

Though body fluid is most readily apparent in the bloodstream, it is present, and in larger amounts, in the tissues, both between the cells (interstitial fluid) and within them (intracellular fluid). Extracellular fluids, which include interstitial fluid and blood plasma, amount to 25 percent of body weight and contain sodium as their predominant cation (positive ion; metals and hydrogen in solution are cations). Intracellular fluids, amounting to 33 percent of body weight, have potassium as their predominant cation. These various "compartments" of body fluid are in osmotic equilibrium, so that if solute (*e.g.*, sodium chloride) is added to the extracellular compartment so as to increase the concentration of the extracellular solution, water will join it to reduce the concentration, and that compartment will increase. An increase in extracellular fluid, if it is considerable, may be clinically apparent as edema, a swelling of the tissues by fluid, which can usually be displaced by firm pressure. Edema is present in acute inflammation of the kidney (nephritis), in protein deficiency of renal origin, and in chronic nephritis complicated by heart failure associated with abnormally high blood pressure; a factor common to all these states is failure of the kidneys to excrete sodium and water in adequate amounts.

Edema and dehydration

The kidneys in such edematous states need not themselves be diseased; for example, normal kidneys, in a patient with heart failure, may retain sodium when handicapped in their function by poor circulation and by abnormal amounts of sodium-retaining hormones, such as aldosterone. Increase in extracellular fluids is the only volume change that is both common and easily discernible in renal disease, but the opposite condition, sodium depletion or clinical dehydration, is more commonly the result of vomiting and diarrhea when they are complications of terminal renal disease. Sodium and water depletion can be recognized by a lack of elasticity in the superficial tissues and by poor filling of the blood vessels, as well as by signs of impaired circulation, including a fall in blood pressure and an increase in pulse rate. Though changes in intracellular fluid volume occur in some diseases, especially when the potassium content of the body is affected, there is no easy way of detecting them.

Properties of body fluids. Because of the importance of osmotic forces in determining fluid distribution within the body, an important attribute of body fluid is its overall osmotic concentration, or osmolality. This depends on the concentration of solutes. While all solutes contribute to osmolality, small particles such as sodium or chloride ions are influential out of all proportion to their weight, and indeed account for over 90 percent of the osmolality of plasma. In the context of renal disease, changes in osmolality depend largely on how the kidney handles water. When the kidney either is incapable of conserving water or is not stimulated by ADH of the pituitary to do so, water is lost from the body, and a state of water depletion develops, characterized by increasing osmolality of body fluid. At other times, the kidneys may retain too much water, especially when too much hormone is present; in this case, water excess results, giving a clinical state of water intoxication, with decreased osmolality of body fluids.

Change in hydrogen-ion concentration

Another important general property of body fluid is its degree of acidity or alkalinity. The kidneys are involved in the excretion of hydrogen ions, and imperfect function leads to their retention, the state of so-called renal acidosis. Renal acidosis may occur as part of general renal failure or as a specific disease of the renal tubules, one of whose functions is to convert the slightly alkaline glomerular filtrate into the (usually) acidic urine.

Apart from these general changes in body fluid, the pattern of individual constituents can be distorted in renal disease. For many substances, the problem is one of failure of excretion, with consequent increased concentration in body fluids. Insofar as excretion is achieved by filtration, the rise in concentration may assist excretion, permitting prolonged states of balance, at the cost of increased, but often tolerable, levels of concentration. For example, an individual in renal failure must put out as much urea as a healthy individual taking the same diet; but that person can only do so at a blood-urea concentration of 100 milligrams per 100 millilitres, instead of a normal blood-urea of 25 milligrams per 100 millilitres. Substances whose concentration increases in this way include urea, creatinine, uric acid, phosphate, sulfate, urochrome, and indeed all the usual constituents of urine apart from those that are "regulated" rather than simply "excreted." Potassium should be mentioned because of the special danger associated with its retention, which can lead to fatal irregularity of cardiac action. This is a recognized danger of acute renal failure, now commonly prevented by use of the artificial kidney and its semipermeable membranes, and sometimes by the use of resins that will take up potassium in the alimentary tract.

Normal urine contains traces of protein, and in many forms of renal disease there is an increased excretion of protein in the urine, usually representing an increased permeability of the tuft of capillaries forming the glomerulus. This increased proteinuria (often, but less correctly, known as albuminuria) generally amounts to 0.5 gram per day or more. When it exceeds five grams per day and persists at this level, the loss of protein in the urine exceeds the capacity of the liver to produce new protein from the available materials; the concentration of protein in the blood decreases, and this leads to an increasing outflow of fluid from the bloodstream into the tissues (there is normally an equilibrium between the physical pressure in the capillaries, which tends to force fluids out, and the osmotic pressure of plasma proteins, the effect of which is to hold fluid in). This balance of forces is upset by a deficit of plasma proteins. The general loss of fluid into the tissues leads to massive edema, to which the kidneys contribute further by retaining salt and water. The combination of high levels of protein in the urine, low protein levels in the blood, and consequent edema is known as the nephrotic syndrome. This is a good example of a syndrome, defined as a recognizable pattern of manifestations that has not one but a number of possible causes. Other examples of syndromes in renal disease are acute renal failure and chronic renal failure.

Disorders of urine flow. If little or no urine appears, it may be because the kidneys are forming little urine (oliguria) or none (anuria); or it may represent a holdup in the bladder or urethra affecting the outflow from both kidneys. About one person in 500 is born with only one kidney, and loss of a kidney from disease or accident is not rare. In such cases, patients with complete obstruction to the remaining ureter will experience the same effect as in obstruction of the entire lower urinary tract. Partial or complete failure to form urine is treated in the section on acute renal failure, obstructive conditions in the section on diseases of the urinary tract. (D.A.K.B./J.S.Ro.)

Damage to nervous control

In instances of damage to nervous control, certain typical clinical situations may be differentiated, corresponding to different modes of disordered urinary flow: (1) Lack of conscious inhibition of micturition because of damage to the cerebral cortex or, more commonly, from psychological causes results in a need to micturate that cannot be suppressed even though the bladder volume may be quite small; micturition is precipitate and continues until the bladder is empty. (2) Transverse lesions or other damage to the spinal cord above the sacral reflex centres that cause paralysis of the lower half of the body produce at first a bladder that is atonic (lacking in physiological tone). This bladder becomes greatly distended; the detrusor relaxes and reflex micturition is abolished. Pressure finally rises sufficiently to overcome the spasm of the sphincters and urine is voided in small amounts. Further accumulation and partial voiding of the overflow recur (overflow incontinence). Under these conditions the bladder readily becomes inflamed, which may cause disability or death from chronic ascending urinary infection. A catheter may be inserted or firm pressure on the lower abdominal wall may be used to avoid overdistension and to develop an "automatic" bladder after some time. This is a small capacity organ (around 150 millilitres) with frequent emptying; there is reflex control mediated through the sacral segments of the spinal cord; the higher centres do not restrain the detrusor, and the internal sphincter relaxes more readily. Voluntary assistance from the abdominal muscles helps in this situation if these too have not been paralyzed. There is, however, always some residual urine from incomplete emptying and a risk of infection. (3) In contrast, there is the isolated, or "autonomous," bladder resulting from damage to the central nervous system below the sacral cord reflex centres or to the nerves supplying the bladder and urethra. The bladder becomes tense but contracts only weakly so that, while small amounts of urine are voided, the residual urine may be as high as 200–300 millilitres. This condition is known as active incontinence as opposed to the overflow incontinence of the automatic bladder. Here again, active support from the abdominal muscles is helpful. (D.LeV./J.S.Ro.)

Dysuria

Pain associated with urination (dysuria) can arise from bladder distension, which is then relieved by effective micturition; from inflammation of the lower urinary tract, commonly due to infection but rarely caused by chemical irritants in the urine; and from mechanical irritation by tumour or during the passage of stones. Dysuria is commonly, but not necessarily, associated with frequency of urination. This in turn may represent either an irritable or contracted bladder; or the actual amount of urine formed may be unusually large (polyuria), in which case voiding is likely to be painless. Sometimes polyuria may not be noticed by day but may manifest itself in the need to micturate on several occasions during the night (nocturia). The acute onset of dysuria and frequency suggests urinary infection; sustained polyuria is more likely to be due to renal failure (defective concentrating power) or to diabetes. In those who drink beverages into the night, nocturia is physiological.

Incontinence of urine

Incontinence, the involuntary passage of urine (or feces), may be due to a faulty nerve supply, which either leaves the sphincters relaxed or allows them to be overcome by distension of the bladder. Comatose and disturbed patients, especially among the elderly, are commonly incontinent. Apart from nerve lesions, the sphincters that normally prevent the escape of urine may be damaged by repeated childbirth, by the growth of the prostate, or by other distortions of the bladder neck. Procedures have been devised to stimulate the sphincters electrically, when their nerve supply is damaged; or to stimulate the bladder to empty itself at set times. For chronic incontinence, however, devices to catch the urine and prevent soiling of clothing are the most practical.

DISEASES AND DISORDERS OF THE KIDNEY

In this section, attention is directed not only to specific diseases of the kidney but also to the syndromes of acute and chronic renal failure, which have multiple causes. Infective disorders of the kidney are dealt with later, as part of the general problem of infection of the urinary tract.

Acute renal failure. Acute renal failure occurs when renal function suddenly declines to very low levels, so that little or no urine is formed, and the substances, including even water, that the kidney normally eliminates are retained in the body. There are two main mechanisms that can produce acute renal failure. When the cardiac output—the amount of blood pumped into the general circulation by the heart—is lowered by hemorrhage or by medical or surgical shock, the renal circulation is depressed to an even greater extent. This leads directly to inefficient excretion, but, more importantly still, the kidney tissue cannot withstand prolonged impairment of its blood supply and undergoes either patchy or massive necrosis (tissue death). Given time, the kidney tissue may regenerate, and it is on this hope that the treatment of acute renal failure is based. The form of acute renal failure that is due to a poor supply of blood (ischemia) has many causes, the most common and most important being multiple injuries, septicemia (infections invading the bloodstream), abortion with abnormal or excessive bleeding from the female genital tract, internal or external hemorrhage, loss of fluid from the body as in severe diarrhea or burns, transfusion reactions, and severe heart attacks; a special case is the transplanted kidney, which commonly goes through a phase of acute renal failure that is independent of possible rejection.

Renal failure due to poisons

The second common mechanism of acute renal failure is toxic. Many poisons are excreted by the kidney, and in the process, like other urinary constituents, they become concentrated and thus reach levels in the tubular fluid that damage the lining cells of the tubules. Though the tubular cells die and are shed in the urine, regeneration can take place and the patient survive, if he can be maintained during the period of depressed renal function and is not killed by other effects of the poison. Poisons that can affect the kidney in this way are numerous, but the main groups are heavy metals (mercury, arsenic, uranium); organic solvents (carbon tetrachloride, propylene glycol, methanol); other organic substances (aniline, phenindione, insecticides); and antibacterial agents (sulfonamides, aminoglycosides, amphotericin), and some fungi (*e.g., Amanita phalloides*). In addition to the ischemic and toxic causes of acute renal failure, mention must be made of fulminating varieties of acute renal illnesses that are generally mild (*e.g.,* acute glomerulonephritis—see below) and of the acute form of immunologic rejection that can destroy a kidney irrevocably within minutes of transplantation. Acute obstruction to the flow of urine from the kidneys obviously also can imperil renal function.

The course of acute renal failure can usefully be divided into three phases: an onset phase, a phase of established acute renal failure, and a recovery phase. In general, but not invariably, the second of these phases is characterized by a low output of urine (oliguria) and the third by an increasing urine output (polyuria). The onset phase is dominated by general illness, in which the episode of acute renal failure arises; at this stage there may be evidence of threatened renal damage such as blood in the urine or pain in the loins. At this early stage, renal damage may be reversible by prompt treatment of circulatory failure (*e.g.,* by the transfusion of adequate amounts of plasma, whole blood, or electrolyte replacement fluids) and by maintaining adequate blood oxygen levels. Infection or any underlying causative disorder also must be treated quickly.

Renal failure: second phase of clinical course

In the second phase, small amounts of urine, often containing red blood cells, or hemoglobin, are passed; complete absence of urine is not common and suggests that an obstruction is preventing urine from being passed. In quantitative terms, a urine volume of less than 500 millilitres per day constitutes significant oliguria; this is the least amount in which the excretory demand imposed by an ordinary diet can be met. In the actual situation of acute renal failure, the excretory demands may in fact be much greater, since many of the causes of acute renal failure also are causes of increased breakdown of the tissues in general. The blood urea increases, the rate of increase being conditioned both by the degree of renal failure and by the amount of tissue breakdown. Besides nitrogen, the kidney can no longer excrete adequate amounts of water, sodium, and potassium.

These various inadequacies point the way to the necessary management of acute renal failure—the elimination from intake of any dangerous substance that the kidney can no longer handle. The diet must either be free of protein or contain small amounts of high-quality protein to lessen tissue breakdown. It must also be free from sodium and potassium: many persons with renal failure have died from pulmonary edema, a correlate of sodium retention, and others from the acute toxic effects on the heart of a raised level of potassium in the blood. Water cannot be excluded from the intake but must be limited to an amount estimated to equal the unavoidable loss of water from the skin and in breathing. The weight of the patient and the concentration of sodium in the blood are good guides to the adequacy of water restriction. In the absence of continuing losses of sodium from the body, as might occur from vomiting or diarrhea, a progressive fall in serum sodium implies that too much water is being taken in. Kidney function may recover, often in seven to 10 days. The use of dialysis, the removal of waste products by straining the blood through semipermeable membranes, gives further time for renal recovery. Potassium can be removed from the body by resins, but this is less often required if dialysis is available.

Although by comparison with the oliguric phase the recovery phase presents fewer problems, the convalescent kidney takes time to recover its full regulatory function, and electrolytes and water may be lost at an unusual rate during this stage, requiring replacement. Most individuals who survive completely recover from acute renal failure, but residual renal damage persists in some persons. In a few, this is so severe as to bring them effectively into the category of chronic renal failure. The artificial kidney has transformed the outlook for many patients with acute renal failure, and this, together with developments in the control of infection with more powerful antibiotics, constitutes one of the miracles of medicine in the last few decades.

Chronic renal failure. The term uremia, though it is sometimes used as if it were interchangeable with chronic renal failure, really means an increase in the concentration of urea in the blood. This can arise in many acute illnesses in which the kidney is not primarily affected and also in the condition of acute renal failure described above. Uremia ought to represent a purely chemical statement, but it is sometimes used to denote a clinical picture, that of severe renal insufficiency.

Chronic renal failure: common causes

As with acute renal failure, there are many conditions that can lead to chronic renal failure. The two most common causes are pyelonephritis and glomerulonephritis (kidney inflammation involving the structures around the renal pelvis or the glomeruli), and other common causes are renal damage from the effects of high blood pressure and renal damage from obstructive conditions of the lower urinary tract. These primary disorders are described below. They have in common a progressive destruction of nephrons, which may be reduced to less than a 20th of their normal number. The quantitative loss of nephrons can account for the majority of the changes observed in chronic renal failure; the failure in excretion is due directly to loss of glomerular filters, and other features such as the large quantities of dilute urine represent a change in tubular function that could be accounted for by the increased load that each remaining nephron has to carry. There are many other causes of chronic renal failure aside from the four common ones. They include congenital anomalies and hereditary disorders; diseases of connective tissue; tuberculosis; the effects of diabetes and other metabolic disorders; and a number of primary disorders of the kidney tubules. Of the many causes, there are some that have importance out of proportion to their frequency, by virtue of their reversibility; these include renal amyloidosis (abnormal deposits in the kidney of a complex protein substance called amyloid), whose causes may be treatable; damage to the kidney from excessive calcium or deficiency of potassium; uric acid deposition in gout; the effects of analgesic agents (substances taken to alleviate pain) and other toxic substances, including drugs.

Effects of chronic renal failure

The person suffering from renal failure, especially in the early stages, may have no symptoms other than a feeling of thirst and a tendency (shared with many normal people) to pass urine at frequent intervals and through the night; or he may be in a coma, with occasional convulsions. The general appearance of the sufferer may be sallow because of a combination of anemia and the retention of urinary pigment. Even if not in actual coma, the affected person may be withdrawn; muscle twitchings and more general convulsions may occur. The coma is thought to represent poisoning, and convulsions are often related to the severity of the high blood pressure that commonly complicates advanced renal failure. Blurred vision is also a manifestation associated with high blood pressure. Bruising and hemorrhages may be noticeable.

Although the toxin (or toxins) of uremia has yet to be identified, the rapid improvement that follows dialysis points strongly to a toxic component. Urea itself is not notably toxic. Not all the chemical alterations in uremia are simple retentions. There is acidosis—a fall in the alkalinity of the blood and tissue fluids—reflected clinically in deep respiration as the lungs strive to eliminate carbon dioxide. The capacity of the kidney to adjust to variation in intake of salt, potassium, and water becomes progressively impaired, so that electrolyte disturbances are common. Poor appetite, nausea, vomiting, and diarrhea are common in uremic patients, and these in turn add another component to the chemical disturbance. Phosphate is retained ,in the blood and is thus associated with low blood levels of calcium; the parathyroids are overactive in renal failure, and vitamin D is less than normally effective because the kidneys manufacture less of its active form (1,25-dihydroxycholecalciferol). (Parathyroid hormone causes release of calcium from the bones, and vitamin D promotes absorption of calcium from the intestines.) These changes can lead to severe bone disease in persons suffering from renal failure, because bone calcium is depleted and the calcium stores are not adequately replenished.

In chronic renal failure, excessive production of renin by the kidney can lead to severe high blood pressure (hypertension), and the effects of this may even dominate the clinical picture. In addition to damage to the brain and the retina, the high blood pressure may lead directly to heart failure. Hypertension can also accelerate the progress of renal damage by its impact on the renal blood vessels themselves, setting up a cycle that can be hard to break. Anemia is also often severe due in part to a failure to produce erythropoietin.

The patient in advanced renal failure is vulnerable to infection and other complications, such as vomiting or diarrhea, which need special care. When symptoms of advanced renal failure appear, deterioration can be delayed by a strict low-protein diet, 18–20 grams of high-quality protein each day. In terminal renal failure, the affected person can be rescued only by some form of dialysis and then maintained by dialysis or transplantation.

Glomerulonephritis. Glomerulonephritis is the disorder commonly known as nephritis, or Bright's disease. The primary impact of the disease is on the vessels of the glomerular tuft. The suffix "-itis" suggests an inflammatory lesion, and glomerulonephritis is indeed associated with infection, in the limited sense that it may begin soon after a streptococcal infection and may be aggravated in its later course by infections of various kinds. Nevertheless, there is convincing evidence that glomerulonephritis does not represent a direct attack on the kidney by an infective agent; it appears to be, rather, an immunologic disorder, in the sense of the formation of antibodies in response to the presence of a foreign protein (antigen) elsewhere in the body; these form antigen–antibody complexes that lodge in the glomerular tuft or, in a small number of cases, themselves become deposited on the capillary glomerular walls. In each case the antibody or the antigen–antibody complex reaches the kidney via the circulation, and the mechanism is usually referred to as circulating complex disease. Glomerular damage is a consequence of the reaction that follows within the glomeruli. These deposits of foreign protein and complexes react with other protein components of blood (see the article COMPLEMENT in the *Micropædia*) and attract to the site white blood cells and platelets, which also are circulating in the blood; these in turn release protease enzymes and other chemical mediators of tissue injury.

Glomerulonephritis: allergic disorder

This view of glomerulonephritis is based partly on analogy with the renal damage that can be induced in animals by allergic mechanisms and partly on finding that a protein component of the allergic reaction is deposited in the diseased glomerulus. Within the general concept of an immunologic disorder, there is ample room for a variety of primary stimuli and of later immunologic disease-causing mechanisms. These include the possibility of primary glomerular damage, causing the glomerulus itself to become antigenic and so to provide a secondary antibody response, and also the participation of (or lack of participation of) T lymphocytes. Such a diversity is strongly suggested not only by the variations in the glomerular tissues observed both with the ordinary and with the electron microscope but also by the varying manifestations of the disease observed in the affected person.

Typically, glomerulonephritis appears as an acute illness one to two weeks after a sore throat, or—less commonly—after a persistent streptococcal infection of the skin. Other infective agents may be responsible, however, including some viruses and protozoans. A small number of drugs that act as foreign macromolecules can also do so.

The affected person has puffiness of the face and ankles and at the same time scanty and noticeably blood-stained urine. On examination, loose tissues show edema, and the fluid is easily displaced by light pressure; both the blood pressure and the blood levels of urea are slightly or moderately increased. The illness is an alarming one, but the fact is that the acute attack of glomerulonephritis needs no particular treatment other than the eradication of the infection or withdrawal of the offending drug, with some restriction of fluid and protein. Nine out of 10 affected persons recover completely. Exceptional outbreaks, with a higher mortality, have sometimes been observed. A very few patients may die in the acute attack, however, or in a few months' time, when the impact of the disease has been unusually severe. Another possibility is that the affected person may appear to have recovered completely, having lost all symptoms; but the disease process remains active, and there is progressive loss of nephrons, leading ultimately to chronic renal failure. This process may take many years, for most of which the person has no definite symptoms of latent nephritis except that the urine contains protein and small numbers of red blood cells. It need not be assumed, however, that the finding of protein in the urine (proteinuria) in the absence of symptoms means automatically that the patient has kidney disease; symptomless proteinuria has many causes and may indeed be found in young people who never develop any later evidence of renal disease.

In summary, glomerulonephritis can lead to renal failure within a few weeks or months, after many years of symptom-free proteinuria, or after a period of massive proteinuria, which causes the nephrotic syndrome. All of these manifestations may sometimes be seen in individuals who have never had, or cannot recall, an acute attack. Renal biopsies in many patients with glomerulonephritis show a range of glomerular reactions that include increased cellularity and basement membrane damage and thickening and varying degrees of progressive destruction of glomeruli. In those who recover, complete resolution of glomerular disease occurs.

A curious form of glomerulonephritis especially common in children is associated with little structural glomerular damage, at least as seen by the ordinary light microscope. Characteristic abnormalities affecting podocytes are revealed by electron microscopy. The condition is usually attended by heavy proteinuria and the nephrotic syndrome. Although the evidence for an immunologic cause of this form of glomerulonephritis is less certain than in other types, and the provoking antigen is unknown, paradoxically the disorder usually promptly resolves when the patient is treated with corticosteroids or other immunosuppressive drugs, and renal failure never occurs.

Vascular disease. In the discussion of chronic renal failure, attention was drawn to the cycle in which high blood pressure secondary to renal disease can produce further damage to the kidneys. Clearly, primary vascular disease—disease affecting the blood vessels—could equally well be a cause of renal damage.

The most dramatic instance of this is the condition known as malignant hypertension, or accelerated hypertension, which arises when the blood pressure attains extremely high levels, the diastolic figure (the blood pressure between heart contractions) being 140 millimetres of mercury or higher (the normal being around 80). Sustained levels of this magnitude cause serious damage to the arterioles, the smallest of the arteries; this damage is widespread, but as it affects the kidneys it produces rapid destruction of renal substance, with a scarred kidney. Unless the blood pressure is controlled, malignant hypertension can cause death in a few months; since treatment at an early stage is notably effective, the condition represents an important medical emergency. Since the retinas are damaged as rapidly as the kidneys, the affected person may first notice blurring or loss of vision.

More modest, but still elevated, levels of blood pressure can cause more gradual renal damage in elderly people or in those made prematurely aged by widespread arteriosclerosis ("hardening of the arteries"). In this condition the damage is in the larger arteries rather than in the arterioles, and the condition is one of slowly progressive scarring. Renal damage can also arise, by various mechanisms, in a large number of diseases that impair the proper functioning of the blood vessels, such as diabetes mellitus, the collagen disorders, bacterial inflammation of the heart lining, and many more.

A specific renovascular cause of high blood pressure that, although uncommon, is important from the point of view of the control of blood pressure in healthy individuals involves the juxtaglomerular apparatus (JGA) and the secretion of renin. Occasionally, following trauma or arising spontaneously as a result of vascular disease, one or the other of the main renal arteries becomes constricted (renal artery stenosis). The fall in blood pressure beyond the constriction leads to increased secretion of renin from the JGA with the formation of the vasoactive angiotensin II. As a result, the blood pressure rises. Removal of the affected kidney, surgical repair of the constriction, or percutaneous transluminal angioplasty (a balloon catheter inserted through the skin and inflated in the artery to flatten plaque build-up) usually restores the blood pressure and the blood renin level to normal.

Tumours. Tumours in general are covered in the article CANCER. In this section, those tumours peculiar to the excretory system, and their local effects, are discussed briefly. In the case of benign tumours, these effects include pressure on local structures and obstruction to hollow organs; with malignant tumours, one must add the possibilities of local invasion and of spread by the bloodstream or lymphatics to other organs (metastasis).

Carcinoma. The most common tumour of the renal substance is a carcinoma, renal cell cancer (formerly called a hypernephroma), which is a malignant tumour, arising from epithelial cells (the cells of the bodily coverings and linings). It was formerly thought to arise from adrenal cortical cells lying within the kidney substance. This has since been disproved. One to 2 percent of all tumours are renal carcinomas, and most affected persons are aged from 40 to 60. The tumour may be symptomless or may first be apparent from the occurrence of metastases in the lungs, causing spitting up of blood; or in the bones, causing pathological fracture.

Signs and symptoms of carcinoma

Much more commonly, the first evidence of the tumour is blood in the urine, which may be painless or may cause colic of the ureter, if clots are being passed. There may also be a dull pain in the loins, from stretching of the kidney capsule. The tumour may be directly palpable, or it may be revealed by X rays or ultrasonography. The silhouette of the kidney may be distorted by a rounded swelling; or the renal pelvis, made visible by the injection of a contrast medium, may be displaced or distorted. Less common first indications of renal carcinoma are an obscure fever, or polycythemia (excess of red blood cells in the blood), due to excessive production of erythropoietin. Direct visual examination of the urinary tract with an instrument called a cystoscope may demonstrate the side that is affected, blood coming from one ureteric opening only. Since this bleeding can equally arise from a tumour of the renal pelvis, examination of the renal pelvis is usually called for. An exploratory operation may sometimes be needed; if carcinoma is found to be present, the kidney must be removed. There is some evidence that the results of surgery may be somewhat improved by radiation therapy. The overall outlook is poor, with a five-year survival rate no better than 50 percent. This is, however, one of the forms of malignant tumour in which arrest or even regression has been described.

Nephroblastoma (Wilms' tumour). Nephroblastoma is a less common, but nevertheless an important, tumour in childhood, in which other forms of cancer are less common. About half the cases occur at ages two to four, but the tumour may be present even at birth. Early diagnosis, immediate surgery, and chemotherapy constitute the best possibility for a cure.

Other tumours. In addition to tumours of the renal substance, the renal pelvis may be affected by fernlike growths of the epithelium (papillomas). Benign tumours of the kidney substance occur, but rarely; on the other hand, cysts (abnormal sacs filled with liquid or semisolid substance) of the kidney are relatively common but are not tumours in any strict sense, being rather malformations brought about by failure of the embryonic tubules to achieve a proper outlet. Several forms of renal cystic disease, most of them fatal, occur in infancy. Various forms of solitary cyst occur, which may need local surgical treatment if they cause symptoms. The form of polycystic (multiple-cyst) renal disease that allows survival into adult life is a familial condition, in which several members of the family have little trouble until middle life but then are progressively affected by kidney malfunction. Episodes of blood in the urine and urinary infection are common, and the kidneys are large and irregular. Cysts of other organs—*e.g.*, the liver—may be present. X rays show irregularity of the renal pelvis, through pressure from the cysts. Puncture of the cysts is possible, but the results are not encouraging; the general treatment is that of chronic renal failure, which may now include removal of the kidney and transplantation.

OBSTRUCTION TO THE FLOW OF URINE

Effect of urinary obstruction on kidney

The causes of obstruction to the flow of urine lie in the lower urinary tract and are dealt with in a later section; here it is appropriate to consider the effects of urinary obstruction on the kidney (obstructive nephropathy). It should first be noted, however, that obstructions may arise at the junction of the renal pelvis and the ureter, either from faulty action of smooth muscle or from the pressure of an abnormal blood vessel crossing the pelvis; such cases can benefit from a plastic operation on the renal pelvis or from division of the abnormal vessel. Whether the obstruction arises in this way, or lower down, it can lead to renal pain, to the passage of irregular amounts of urine when obstruction is intermittent, and to a mass in the kidney when obstruction persists. As the renal pelvis swells, the renal tissue shrinks, leading to the condition called hydronephrosis, in which a greatly swollen sac is surrounded by a mere rind of atrophied renal tissue. A massive hydronephrosis, with negligible renal substance remaining, may suggest removal of the kidney.

Trauma

The kidney may be wounded, usually along with other viscera; it may be bruised; or it may even be ruptured in closed injuries. Since the kidney receives about a fifth of the blood pumped by the heart, bleeding can be profuse, both into the urine and into the tissues and the kidney, forming a large mass of blood, called a hematoma, and leading to surgical shock. Some bleeding may follow the procedure of renal biopsy (taking a specimen of kidney tissue for examination), but with proper precautions this is not severe. In the past, massive irradiation to the kidney region led to chronic renal damage (radiation nephritis), but with adequate precautions, this is no longer so.

The usual signs of traumatic injury to the kidney are blood in the urine and the development of a tender mass in the loin, with progressive signs of shock (pallor, sweating, fall in blood pressure). Such signs call for resuscitation and for surgical exploration if the bleeding continues. The surgical treatment may be carried out to arrest the bleeding by closing the tear. The kidney must be surgically removed if it cannot be saved. Abnormal solitary kidneys are not unknown, and such kidneys are more exposed to trauma by their size or position. Removal of such a kidney can lead only to death unless transplantation is possible.

SUBSTITUTING FOR RENAL FUNCTION

The failure of a vital function normally, and by definition, leads to death; but in the case of the kidneys there are two methods of substituting for renal function: transplantation and dialysis.

Transplantation. In principle the simpler of these two is to transplant a kidney from a donor, ideally an identical twin. The immunologic and surgical problems of transplantation are dealt with in the article TRANSPLANTS, ORGAN AND TISSUE. Here only the part played by renal transplantation in the total care of renal disease is considered. The question of a transplant does not arise in most cases of acute renal failure when the loss of function is largely recoverable; and in chronic renal failure it arises only when the residual renal function is barely adequate to support life.

Occasions for use of transplants and dialysis

Dialysis. Transplantation and dialysis are complementary rather than rival methods. Dialysis is used while a patient is awaiting transplant and during episodes of oliguria or of threatened rejection, while, on the other hand, patients who find dialysis a psychological burden can be offered a transplant. In addition to its complementary role in a transplant program, dialysis can be used independently in the maintenance of patients with chronic renal failure; and it can be used to preserve life in acute renal failure and in acute poisoning, to allow more time for recovery.

There are two main techniques of dialysis in current use. In peritoneal dialysis, the patient's own abdominal cavity is used as the container of fluid; the fluid is run in, allowed to reach equilibrium, and removed, taking with it urea and other wastes. The process has proved suitable for the short-term treatment of acute renal failure, especially in infants, and can be used in the treatment of individuals with chronic irreversible renal failure. New techniques have allowed many patients to conduct peritoneal dialysis on their own for limited periods of time.

Hemodialysis (filtration of the blood through semipermeable membranes) has also been used in the treatment of acute renal failure, since the method—the artificial kidney—was devised, in the 1940s; but, for chronic use, the problem was one of repeated access to the arterial bloodstream. This was largely solved by the introduction of a permanent shunt between an artery and a vein (an arteriovenous fistula), by which a suitable vein, usually in the arm, is connected directly to an adjacent artery. The vein becomes distended and so can be repeatedly punctured to gain access to blood, which can then be diverted through the "artificial kidney" when required. In the original artificial kidney, the patient's blood was pumped through cellophane tubing immersed in a large bath of physiological fluid (solution of the same osmotic pressure as blood); in some later models, streams of blood and of dialyzing fluid are made to flow in opposite directions, separated by plastic sheets. This introduction of the "countercurrent" principle has allowed the apparatus to be smaller, and disposable versions of both patterns are now available. Some patients on intermittent hemodialysis have been kept alive for nearly 20 years. Most continued hemodialysis is still done in hospitals or special centres; but some patients using automatic equipment have been successfully trained to carry out the procedure in their own home.

DISEASES AND DISORDERS OF THE URINARY TRACT

Obstruction. While it is possible for the urinary tract to be obstructed by a large mass (tumour, stone, or foreign body) lying in the bladder, the tubular portions of the tract (urethra and ureters) are much more vulnerable to obstruction. The urethra may be obstructed by stones (calculi) formed in the bladder or kidneys; by fibrous contraction of the urethral wall (urethral stricture); and by congenital valve or diaphragm (membranous malformation). Although not a part of the excretory tract, the prostate lies close to the bladder neck, and in older men it is an important cause of obstruction; fibrous disease of the bladder neck can also cause obstruction. The ureters can likewise be obstructed by calculi and stricture (narrowing); by fibrosis—scarring—of surrounding tissue (retroperitoneal fibrosis); and by tumour, though this is more likely to cause blood in the urine (hematuria).

Characteristics of urinary calculi

Urinary calculi vary greatly in size. Mostly they contain calcium phosphate, calcium oxalate, uric acid, or cystine. Predisposing factors include infection, a high rate of calcium excretion, a low rate of urine formation, and various metabolic disorders, notably gout. They may cause trouble by their size or by entering the ureter or urethra, giving rise to colic, to hematuria, and, in the event of impaction, to obstructive kidney disease. The direct treatment of calculi is surgical, but sometimes the stone can be fragmented in situ by a lithotriptor. The sufferer needs general investigation for any underlying cause (*e.g.,* a functioning parathyroid tumour that causes excessive excretion of calcium).

In the past at least, a common cause of urethral stricture was gonorrhea, in which inflammation of the urethra is followed by scarring and stricture. Bruising of the urethra by instruments during treatment can also occur. The affected person has increasing difficulty in passing urine, and the bladder becomes distended. Treatment may be either by repeated dilation of the stricture or by surgery.

Trauma. Apart from the urethra, the urinary tract is likely to be injured only in massive general injury or by accidental ligation (tying) of the ureters in a pelvic operation. The urethra can, however, be ruptured by a blow or fall on the perineum (crotch). If there is no external wound, the damage is indicated by the appearance of a swelling containing blood and urine, by the inability to pass urine, and by bleeding from the urethra. The patient becomes shocked and urgently needs surgical repair of the urethra and drainage of the potentially infected swelling.

Tumour. The occurrence of papillomatous tumours of the renal pelvis has already been mentioned. Similar tumours in the lower urinary tract give rise to painless hematuria. Workers with the chemicals naphthylamine and benzidine have a high incidence of bladder tumours, often multiple and recurrent. Blood in the urine is the most frequent symptom, but bladder irritation with difficulty in urination appears later. Removal when practicable or destruction by diathermy are normal treatments.

Situations leading to urinary infection

Infection of urinary tract. Infection of the urinary tract is a common and important cause of both minor and major illness. At one extreme, an attack of cystitis—inflammation of the bladder—may cause only trivial discomfort; on the other hand, infection once established may cause lifelong discomfort, may be largely unresponsive to treatment, and may greatly shorten life itself. Infection may be with a great variety of organisms, but the most common are those that normally inhabit the bowel, where they are relatively harmless, becoming a cause of disease only when they enter vulnerable tissue. Because of the short female urethra, urinary infections are more common in women than in men and occur especially during pregnancies, when there may be partial stagnation of the urine from pressure on the urinary tract. In later life, as prostatic disease becomes more common, urinary infection becomes more of a problem in men. Another vulnerable period is infancy, when the use of diapers probably facilitates entry of organisms into the urethra. The introduction of a catheter into the bladder may be necessary to relieve urethral obstruction, but since the procedure always carries a risk of introducing infection, it is not lightly undertaken.

In all forms of urinary infection the urine may be cloudy and may contain more ammonia than usual. Urination tends to be painful if the urethra is inflamed, and both painful and frequent if inflammation involves the bladder. Bladder infection may also cause fever, dull pain in the lower part of the abdomen, and vomiting. If the infection reaches the kidneys, symptoms are even more severe, and there is pain in the loins, on one or both sides, and sometimes high fever.

Urinary infection may generally be diagnosed from the symptoms and from laboratory examination of the urine. The treatment is usually the administration of sulfonamides or broad-spectrum antibiotics. The extent to which repeated, or recurrent, urinary tract infection may lead to chronic pyelonephritis (inflammation of the kidney and lining of the renal pelvis) and renal failure remains a controversial issue. It is agreed that, in the presence of obstruction to the flow of urine, urinary infection is prone to ascend the urinary tract and cause intractable infection within the renal pelvis and kidney tissue. Infection can rarely be eradicated by antibiotics until the obstruction is removed or relieved. Although many patients have signs of progressive renal damage they have sterile urine and no signs of infection. Investigations, including direct histological examination of the kidneys, however, reveal that chronic inflammation has been present for many years within and between the renal tubules (interstitial nephritis). Some of these patients admit to excessive and prolonged use of nonsteroidal analgesic drugs such as phenacetin. In others it is possible that urinary tract infection and renal damage began in infancy, possibly encouraged by regurgitation of urine into the ureter and pelvis as a result of an incompetent ureterovesical valve (vesicoureteric reflux). This process not only damages the kidneys directly at an early age but favours the development of infection and leads in later life to the development of kidneys distorted by fibrosis and scar tissue. In any event, pyelonephritis and glomerulonephritis are by far the two most common causes of chronic renal failure sufficiently severe to necessitate dialysis or renal transplantation.

Like other tissues, the excretory system can be involved in tuberculous infection. This is now relatively uncommon and, when it occurs, can often be managed by the general chemotherapy appropriate to tuberculous infection. Advanced renal tuberculosis requiring removal of the kidney rarely occurs. (D.A.K.B./J.S.Ro.)

RENAL DISORDERS IN PREGNANCY

The pregnant woman is especially vulnerable to two renal disorders: acute urinary tract infection and preeclampsia. Acute urinary tract infection, as discussed above, is the most common complication of pregnancy; although it is responsible for much discomfort and distress, it does not affect mortality of either mother or fetus.

While elevation of blood pressure may accompany the onset of pregnancy, the development of rising levels of blood pressure in the last three months of pregnancy is particularly ominous and heralds the onset of a condition known as preeclampsia; this is especially prone to occur in a first pregnancy. In addition to high blood pressure, there is rapid weight gain, fluid retention, and proteinuria. The condition has been described as a "disease of theories," because its cause remains obscure. Its development, however, is certainly linked to the presence of the placenta and fetus within the uterus (womb). It seems likely that an initiating event is insufficient blood flow to the uterus, which in turn leads to ischemia of the placenta; *i.e.*, parts of the placental tissue undergo degeneration or die. This in turn releases substances into the bloodstream that increase the tendency for the blood to clot within renal capillaries and small blood vessels elsewhere in the body. Renal failure and other organ damage then ensues, and hypertension becomes more severe. If the condition is untreated, generalized seizures and convulsions follow (eclampsia). Eclampsia is a serious condition with high fetal and maternal death. It does not develop if preeclampsia is treated sufficiently early. Hypertension must be controlled through drug therapy, and it is desirable that the baby be delivered some weeks before full term. (J.S.Ro.)

BIBLIOGRAPHY

Elimination: PAUL B. WEISZ and RICHARD N. KEOGH, *The Science of Biology,* 5th ed. (1982), a comprehensive general text emphasizing molecular biology, ecology, and morphology; RALPH BUCHSBAUM, *Animals Without Backbones,* 3rd ed. (1987), an elementary, illustrated account of invertebrate animals; ALFRED SHERWOOD ROMER and THOMAS S. PARSONS, *The Vertebrate Body,* 6th ed. (1986), a general history of the vertebrate body emphasizing comparative aspects of structure and function; and C. LADD PROSSER (ed.), *Comparative Animal Physiology,* 3rd ed. (1973), a college-level text on the comparative aspects of functional systems in animals. See also ALBERTE PULLMAN, V. VASILESCU, and L. PACKER (eds.), *Water and Ions in Biological Systems* (1985); and ALBERT BÄR and GÜNTHER RITZEL (eds.), *Xylitol and Oxalate: Metabolic Studies in Animals and Man* (1985).

Excretion and excretory systems: HOMER W. SMITH, *From Fish to Philosopher* (1953, reissued 1961), is an introduction to vertebrate evolution in terms of kidney evolution. Among sources of information on the comparative structure and function of excretory systems are the following: ERNST FLOREY, *An Introduction to General and Comparative Animal Physiology* (1966); EDWIN S. GOODRICH, *Studies on the Structure and Development of Vertebrates* (1930, reprinted 1986), an older work of unchallenged authority on morphological and evolutionary aspects of all vertebrate systems, including the urogenital; MALCOLM S. GORDON *et al., Animal Physiology: Principles and Adaptations,* 4th ed. (1982), on the regulation of water and salts in animals; and WILLIAM S. HOAR, *General and Comparative Physiology,* 3rd ed. (1983), an established authoritative work. Appropriate sections of the multivolume series *Handbook of Physiology,* edited by JOHN FIELD, provide surveys of the excretory function but presume a greater background knowledge than do the works of Hoar and Florey, listed above. See also A. WESSING (ed.), *Excretion* (1975), a collection of symposium papers with bibliographies.

The human excretory system: DON W. FAWCETT, *A Textbook of Histology,* 11th ed. (1986), an illustrated text on histology and fine structure; D.J. CUNNINGHAM, *Cunningham's Textbook of Anatomy,* 12th ed., edited by G.J. ROMANES (1981), a traditional text on human anatomy; HENRY GRAY, *Anatomy of the Human Body,* 30th American ed., edited by CARMINE D. CLEMENTE (1985), a treatise frequently revised over the past century and still regarded as one of the great anatomical texts; ARTHUR W. HAM, *Ham's Histology,* 9th ed. (1987), a text paying particular attention to the correlation of structure and function; and W.J. HAMILTON and H.W. MOSSMAN, *Human Embryology: Prenatal Development of Form and Function,* 4th ed. (1972), one of the leading textbooks on human embryology. D. BULMER, *Functional Anatomy of the Urogenital System* (1974); and JOHN A. GOSLING, JOHN S. DIXON, and JOHN R. HUMPHERSON, *Functional Anatomy of the Urinary Tract* (1982), are specialized texts.

Human excretion: The following works provide detailed and well-written surveys of the field of kidney function in a form suitable for medical students and for others with a good scientific grounding: E.J. MORAN CAMPBELL *et al.* (eds.), *Clinical Physiology,* 5th ed. (1984); and J.M. FORRESTER *et al.* (eds.), *A Companion to Medical Studies: Anatomy, Biochemistry, and Physiology,* 3rd ed. (1985).

Excretory system diseases: For further information, the following may be consulted: H.E. DE WARDENER, *The Kidney: An Outline of Normal and Abnormal Function,* 5th ed. (1985), a clearly written textbook; SIR D.A.K. BLACK and N.F. JONES (eds.), *Renal Disease,* 4th ed. (1979), a multiple-author book that concentrates on selected areas of current development; BARRY M. BRENNER and J. MICHAEL LAZARUS (eds.), *Acute Renal Failure* (1983); BARRY M. BRENNER and FLOYD C. RECTOR, JR. (eds.), *The Kidney,* 3rd ed., 2 vol. (1986); ROBERT W. SCHRIER (ed.), *Renal and Electrolyte Disorders,* 3rd ed. (1986); STUART L. STANTON (ed.), *Clinical Gynecologic Urology* (1984); DONALD W. SELDIN and GERHARD GIEBISCH (eds.), *The Kidney: Physiology and Pathophysiology* (1985); NORMAN M. KAPLAN, BARRY M. BRENNER, and JOHN H. LARAGH (eds.), *The Kidney in Hypertension* (1987).

(F.C.Ke./J.A.R./G.A.G.M./D.LeV./
D.A.K.B./J.S.Ro.)

Exercise and Physical Conditioning

The terms exercise and physical activity are often used interchangeably, but this article will distinguish between them. Physical activity is an inclusive term that refers to any expenditure of energy brought about by bodily movement via the skeletal muscles; as such, it includes the complete spectrum of activity from very low resting levels to maximal exertion. Exercise is a component of physical activity. The distinguishing characteristic of exercise is that it is a structured activity specifically planned to develop and maintain physical fitness. Physical conditioning refers to the development of physical fitness through the adaptation of the body and its various systems to an exercise program.

This article is divided into the following sections:

A HISTORICAL VIEW OF EXERCISE

Prehistoric period. Hominids—human beings and their immediate ancestors—have existed on Earth for at least 2,000,000 years. For more than 99 percent of that time, hominids lived a nomadic existence and survived by hunting and gathering food. It is obvious that this way of life was enormously different from the way people live today in developed countries. Thus, evolutionary history has prepared humankind for one kind of life, but modern people lead another. This fact has profound implications for patterns of disease and for the association between living habits and health. Observation of the few remaining nomadic groups in the world indicates that they are relatively free of chronic diseases and that, in comparison to the populations in developed countries, they are leaner, have a higher level of physical fitness, eat a very different diet, and have different physical activity patterns. Data from the distant past are not available, but it is reasonable to speculate that early humans had considerably higher caloric expenditures per unit of body weight than do modern individuals.

Agricultural period. As civilization developed, nomadic hunting and gathering societies gave way to agricultural ones in which people grew their own food and domesticated animals. This development occurred relatively recently, approximately 10,000 years ago. Although many aspects of life changed during the agricultural period, it is likely that energy demands remained high, with much of the work still done by human power. Even in cities—which had evolved by about midway through the agricultural period—individuals expended more calories than do most people today.

Industrial period. The industrial period began during the mid-18th century, with the development of an efficient steam engine, and lasted to the end of World War II (1945). This relatively brief time span was characterized by a major shift in population from farms to cities, with attendant changes in many areas of life-style. Even though the internal-combustion engine and electrical power were increasingly used to perform work, the great majority of individuals in industrialized societies still faced significant energy demands. In the cities relatively more individuals walked to work, climbed stairs, and had more physically demanding jobs than do most people today.

Effects of labour-saving devices

Technological period. The post-World-War-II period has been a technological age, a period characterized by rapid growth in energy-saving devices, both in the home and at the workplace. As an example, longshoremen in the late 1940s worked hard loading and unloading ships; by contrast, most longshoremen in the late 20th century had much lower energy demands from the job, because of the containerization of cargo and the mechanization of the loading and unloading process. Also during this period, the use of labour-saving devices in the home and in yard and garden work became much more widespread. Physical activity became less and less common in industrialized countries, especially among the urban population. Although the level of general physical activity has declined, most observers feel that there have been increases in exercise participation in many countries since the late 1960s. Jogging, racket sports, cycling, and other active recreational pursuits have become much more common. In a sense this is simply humankind's returning to the more active life-style of its distant ancestors.

TYPES OF PHYSICAL FITNESS

Physical fitness is a general concept and is defined in many ways by different scientists. Physical fitness is discussed here in two major categories: health-related physical fitness and motor-performance physical fitness. Despite some overlap between these classifications, there are major differences, as described below.

Health-related physical fitness. Health-related physical fitness is defined as fitness related to some aspect of health. This type of physical fitness is primarily influenced by an individual's exercise habits; thus, it is a dynamic state and may change. Physical characteristics that constitute health-related physical fitness include strength and endurance of skeletal muscles, joint flexibility, body composition, and cardiorespiratory endurance. All these attributes change in response to appropriate physical conditioning programs, and all are related to health.

Strength and endurance of skeletal muscles of the trunk help maintain correct posture and prevent such problems as low back pain. Minimal levels of muscular strength and endurance are needed for routine tasks of living, such as carrying bags of groceries or picking up a young child. Individuals with very low levels of muscular strength and endurance are limited in the performance of routine tasks and have to lead a restricted life. Such limitations are perhaps only indirectly related to health, but individuals who cannot pick up and hug a grandchild or must struggle to get up from a soft chair surely have a lower quality of life than that enjoyed by their fitter peers.

Flexibility, or range of motion around the joints, also ranks as an important component of health-related fitness. Lack of flexibility in the lower back and posterior thigh is thought to contribute to low back pain. Extreme lack of flexibility also has a deleterious effect on the quality of life by limiting performance.

Body composition refers to the ratio between fat and lean tissue in the body. Excess body fat is clearly related to several health problems, including cardiovascular disease, type II (adult-onset) diabetes mellitus, and certain forms of cancer. Body composition is affected by diet, but exercise habits play a crucial role in preventing obesity and maintaining acceptable levels of body fat.

Cardiorespiratory endurance, or aerobic fitness, is probably what most people identify as physical fitness. Aerobic fitness refers to the integrated functional capacity of the heart, lungs, vascular system, and skeletal muscles to expend energy. The basic activity that underlies this type of fitness is aerobic metabolism in the muscle cell, a process in which oxygen is combined with a fuel source (fats or carbohydrates) to release energy and produce carbon dioxide and water. The energy is used by the muscle to contract, thereby exerting force that can be used for movement. For the aerobic reaction to take place, the cardiorespiratory system (*i.e.*, the circulatory and pulmonary systems) must constantly supply oxygen and fuel to the muscle cell and remove carbon dioxide from it. The maximal rate at which aerobic metabolism can occur is thus determined by the functional capacity of the cardiorespiratory system and is measured in the laboratory as maximal oxygen intake. As will be discussed in detail below, aerobic fitness is inversely related to the incidence of coronary heart disease and hypertension.

Motor-performance physical fitness. Motor-performance fitness is defined as the ability of the neuromuscular system to perform specific tasks. Test items used to assess motor-performance fitness include chin-ups, sit-ups, the 50-yard dash, the standing long jump, and the shuttle run (a timed run in which the participant dashes back and forth between two points). The primary physical characteristics measured by these tests are the strength and endurance of the skeletal muscles and the speed or power of the legs. These traits are important for success in many types of athletics. Muscular strength and endurance are also related to some aspects of health, as stated above.

There is disagreement among experts about the relative importance of health-related and motor-performance physical fitness. While both types of fitness are obviously desirable, their relative values should be determined by an individual's personal fitness objectives. If success in athletic events is of primary importance, motor-performance fitness should be emphasized. If concern about health is paramount, health-related fitness should be the focus. Different types of fitness may be important not only to different individuals but also to the same individual at different times. The 16-year-old competing on a school athletic team is likely to focus on motor performance. The typical middle-aged individual is not as likely to be concerned about athletic success, emphasizing instead health and appearance. One further point should be made: to a great extent, motor-performance physical fitness is determined by genetic potential. The person who can run fast at 10 years of age will be fast at age 17; although training may enhance racing performance, it will not appreciably change the individual's genetically determined running speed. On the other hand, characteristics of health-related physical fitness, while also partly determined by inheritance, are much more profoundly influenced by exercise habits.

The role of
heredity

PRINCIPLES OF EXERCISE TRAINING

Research in exercise training has led to the recognition of a number of general principles of conditioning. These principles must be applied to the development of a successful exercise program.

Specificity. The principle of specificity derives from the observation that the adaptation of the body or change in physical fitness is specific to the type of training undertaken. Quite simply this means that if a fitness objective is to increase flexibility, then flexibility training must be used. If one desires to develop strength, resistance or strengthening exercises must be employed. This principle is indeed simple; however, it is frequently ignored. Many fraudulent claims for an exercise product or system promise overall physical fitness from one simple training technique. A person should be suspicious of such claims and should consider whether or not the exercise training recommended is the type that will produce the specific changes desired.

Overload. Overload, the second important principle, means that to improve any aspect of physical fitness the individual must continually increase the demands placed on the appropriate body systems. For example, to develop strength, progressively heavier objects must be lifted. Overload in running programs is achieved by running longer distances or by increasing the speed.

Progression. Individuals frequently make the mistake of attempting too rapid a fitness change. A classic example is that of the middle-aged man or woman who has done no exercise for 20 years and suddenly begins a vigorous training program. The result of such activity is frequently an injury or, at the least, stiffness and soreness. There are no hard-and-fast rules on how rapidly one should progress to a higher level of activity. The individual's subjective impression of whether or not the body seems to be able to tolerate increased training serves as a good guide. In general it might be reasonable not to progress to higher levels of activity more often than every one or two weeks.

Warm-up/cool down. Another important practice to follow in an exercise program is to gradually start the exercise session and gradually taper off at the end. The warm-up allows various body systems to adjust to increased metabolic demands. The heart rate increases, blood flow increases, and muscle temperatures rise. Warming up is certainly a more comfortable way to begin an exercise session and is probably safer. Progressively more vigorous exercises or a gradual increase in walking speed are good ways to warm up. It is equally important to cool down—that is, to gradually reduce exercise intensity—at the end of each session. The abrupt cessation of vigorous exercise may cause blood to pool in the legs, which can cause fainting or, more seriously, can sometimes precipitate cardiac complications. Slow walking and stretching for five minutes at the end of an exercise session is therefore a good practice. The heart rate should gradually decline during the cool down, and by the end of the five minutes it should be less than 120 beats per minute for individuals under 50 years of age and less than 100 beats per minute for those over 50.

Frequency, intensity, and duration. To provide guidance on how much exercise an individual should do, exercise physiologists have developed equations based on research. It is generally agreed that to develop and maintain physical fitness, the exercise must be performed on a regular basis. A frequency of about every other day or three days per week appears minimally sufficient. Many individuals exercise more frequently than this, and, of course, such additional exercise is acceptable provided that one does not become overtrained and suffer illness or injury.

The intensity of exercise required to produce benefits has been the subject of much study. Many people have the impression that exercise is not doing any good unless it hurts. This is simply not true. Regular exercise at 45 to 50 percent of one's maximal capacity is adequate to improve one's physiological functioning and overall health. This level of intensity is generally comfortable for most individuals. A reliable way to gauge exercise intensity is to measure the heart rate during exercise. An exercise heart rate that is 65 percent of a person's maximal heart rate corresponds to approximately 50 percent of his maximal capacity. Maximal heart rate can be estimated by subtracting one's age in years from 220 (or, in the case of active males, by subtracting half of one's age from 205). Thus, a sedentary 40-year-old man has an estimated maximal heart rate of 180 beats per minute. Sixty-five percent of this maximal rate is 117 beats per minute; thus by exercising at 117 beats per minute, this individual is working at about 50 percent of his maximal capacity. To determine exercising heart rate, a person should exercise for several minutes, to allow the heart rate to adjust. The exerciser should then stop exercising, quickly find the pulse, and count the number of beats for 15 seconds. Multiplying this by four gives the rate in beats per minute. The pulse

Heart rate
as a gauge
of exercise
intensity

must be taken immediately after stopping exercise, since the heart rate rapidly begins to return to the resting level after work has been stopped. As noted above, exercising at the 50 percent level of intensity will improve physiologic functioning and provide health benefits. This level of exercise will not produce the maximum fitness needed for competitive athletics.

There is a relationship between the recommended duration of exercise and the intensity of exercise. If one exercises at a very high intensity, the duration of the exercise need not be as long to produce physiological changes. At lower intensities, duration must increase. The table presents guidelines for reasonable intensity, duration, and frequency.

Guidelines for Intensity, Duration, and Frequency of Exercise

exercise heart rate (beats/minute)	duration of session (minutes)	sessions per week
140	20	4
130	30	3
120	45	5

These values apply to a 40-year-old person.

Overall conditioning. Much emphasis has been given in the foregoing discussion to aerobic fitness, because this form of conditioning is extremely important. It should be noted, however, that other types of conditioning also have benefits. A total exercise program should include strengthening exercises, to maintain body mass and appropriate levels of strength for daily functioning, and stretching exercises to maintain joint mobility and flexibility. The specificity principle described above indicates that no one exercise is likely to produce the overall conditioning effect. In general an exercise plan should consist of aerobics, exercises that increase the strength and endurance of various skeletal muscle groups, and flexibility exercises to maintain good joint function.

Individual differences. The principles of exercise training discussed above should be viewed as general guidelines. Individuals differ in both physiological and psychological adaptations to exercise. Two people who are similar in many respects and who start the same exercise program may have entirely different impressions of it. One person may feel that the exercise is too easy, while the other may believe that it is much too hard. It is certainly appropriate that the exercise plan be adjusted to account for preferences. Likewise some individuals will progress to more intense training levels far more rapidly than others do. As mentioned earlier, exercise progress should be adjusted according to the exerciser's own assessment.

Individuals also differ in the type of exercise they like or can tolerate. Jogging, for instance, is not for everyone. Many people who dislike jogging, or who suffer running injuries, can find other satisfactory exercise activities, such as cycling, walking, swimming, or participating in a sport. Many kinds of exercise activities are appropriate and can provide physiological and health benefits to the participant. There is no one best exercise. The important thing is to be regular in exercise participation and to follow the general guidelines outlined in this section.

PHYSIOLOGICAL EFFECTS OF EXERCISE

Strength training

Neuromuscular effects. *Strength and endurance.* Appropriate exercise increases the strength and endurance of skeletal muscles. Increases in muscular strength are associated with increases in muscle mass; increases in muscular endurance are associated with improved blood flow to the working muscles. These results are achieved by resistance training. Any exercise that causes the muscle to increase its tension, whether or not the muscle actually shortens during contraction, provides an appropriate strength-training stimulus. Resistance can be applied to a muscle group by attempting to move an immovable object, by working one muscle group against another, by lifting heavy weights, or by using special strength-training machines and devices. There is a wide selection of strength-training equipment that, when used properly, can increase muscular strength and endurance. It is possible that some of the equipment is more efficient in developing maximal performance, which is important for competitive athletes. But for the average individual, who is training to maintain an acceptable level of muscular fitness, any one device or program is probably about as good as another.

Strength and endurance training is done by performing several "reps" (repetitions) of a given exercise, then moving on to another exercise for a different muscle group. Experts generally recommend that exercisers select a resistance that is approximately 65 percent of the maximum they can lift for that particular exercise. This load should allow the completion of 12 reps of that exercise in 24 to 30 seconds. Each group of eight to 12 reps is called a set, and two or three sets of a given exercise are recommended for each training session. The average individual should perform strength and endurance training two to three days per week. Super circuit weight training refers to a program in which running or other aerobic exercises are performed between sets; this training produces aerobic as well as strength benefits.

Stretching technique

Flexibility. Muscles and tendons can be stretched to improve flexibility (the range of motion at a joint). Flexibility training follows a few, simple principles. To improve range of motion, the muscles and other connective tissue around a joint must be stretched. The preferred stretching technique is a slow increase in the range of motion. The exerciser should feel the muscle stretch, but not to the point of pain. The stretch should be performed gradually, and the body should be held for 10 to 20 seconds in the stretched position and then gradually returned to a relaxed posture. By stretching each muscle group in this fashion as a part of the strengthening and conditioning program, the participant will maintain good flexibility. Bouncing or explosive stretching movements should be avoided, as they can result in muscle or tendon tears.

Cardiorespiratory effects. *Cardiac effects.* Regular aerobic exercise training has a direct effect on the heart muscle. The muscle mass of the left ventricle, which is the pumping chamber that circulates blood throughout the body, increases with exercise training. This change means that the heart can pump more blood with each beat. In short, the heart becomes a bigger, stronger, and more efficient pump capable of doing more work with less effort.

Circulatory effects. Regular exercise also produces changes in the circulation. As previously discussed, muscle endurance training serves to increase blood flow to the working muscles. This increased blood flow means that more oxygen and fuel can be delivered to the muscle cells. The number of red blood cells, which carry oxygen in the blood, also increases with training, as does blood volume. Taken together, these changes indicate a greater capacity to transport oxygen to the working muscles.

Pulmonary effects. The basic function of the lungs is to facilitate the transfer (1) of oxygen from the atmosphere into the blood and (2) of carbon dioxide from the blood into the atmosphere. To accomplish this, air must pass into and out of the lungs, and the respiratory gases must diffuse through the lungs into the circulation and vice versa. Although exercise has not been shown to affect this diffusing ability, exercise training does strengthen the muscles of respiration. This means that a trained individual can move more air through the lungs per time unit, and forced vital capacity (*i.e.*, the maximum volume of air that can be exhaled after a full inspiration) may be increased.

HEALTH EFFECTS OF EXERCISE

Effects on energy-spending ability

Improved general fitness. The greatest benefit of a regular exercise program is an improvement in overall fitness. As discussed above, appropriate exercise improves muscular strength and endurance, body composition, flexibility, and cardiorespiratory endurance. The level of maximal oxygen intake or cardiorespiratory endurance is not by itself of great importance to most individuals. What is important is that one's sustained energy-spending ability is directly related to maximal levels of performance. For example, consider the simple task of walking at a rate of three miles per hour. This task involves an energy expenditure of approximately three times the resting metabolic rate. Extremely unfit individuals may have a maximal aerobic power of only six times their resting metabolic rate. For such individuals, a three-mile-per-hour walk requires half of their maximal capacity. A middle-aged person who exercises regularly will have a maximal aerobic power 10 to 12 times resting, so the same walk will represent only 25 to 30 percent of maximal capacity. This example illustrates how any submaximal task is relatively much easier for the conditioned individual. Moreover, a person cannot work throughout the day at much more than about 20 percent of maximal capacity without becoming chronically fatigued. The deconditioned person who has a maximal aerobic power of six times resting can comfortably sustain a work level of only about 1.2 times resting throughout the day (6 × 0.20 = 1.2). This low capability for sustained energy expenditure can support only a very sedentary existence: for example, 20 hours of sleep and rest, two hours of personal care, one hour of housework and shopping, and one hour of activity at three times the resting rate each day.

The point of the preceding discussion is that the average energy-expenditure requirement of anyone's life can be calculated, and a person's maximal cardiorespiratory endurance determines how active a life-style can be sustained. Individuals who attempt to lead more active lives than their fitness level will support become chronically fatigued. Persons with adequate or optimal fitness levels, on the other hand, are able to meet the physical demands of an active life relatively easily. One of the most frequent observations made by individuals who have begun an exercise program is that they feel better, and research studies document an improvement in feelings of general well-being in more active people.

Decreased risk of coronary heart disease. Coronary heart disease is the leading cause of death in the developed world. Coronary heart disease is defined as myocardial infarction, or heart attack; angina pectoris, or chest pain; or sudden death due to cardiac arrest or abnormal electrical activity in the heart. The basic disease process that underlies coronary heart disease is atherosclerosis, a disorder characterized by the accumulation of cholesterol and the proliferation of smooth muscle cells in the linings of the arteries. This results in a gradual narrowing of the arterial channel, and this narrowing diminishes and may ultimately stop blood flow through an artery. When this occurs in a coronary artery—that is, an artery supplying the heart—one of the manifestations of coronary heart disease occurs.

Atherosclerosis

Epidemiological evidence of exercise benefits. Studies have linked sedentary living with high rates of coronary heart disease mortality. One study found that San Francisco longshoremen who worked in jobs requiring high levels of energy expenditure had less risk of dying of heart disease than did longshoremen who performed sedentary jobs. This study showed that dockworkers and cargo handlers expended at least 1,000 kilocalories more per day than did clerks and foremen and that sedentary workers, during a 22-year observation, were about twice as likely to die from heart disease. The higher risk of death in the less active men was not due to other coronary heart disease risk factors, such as smoking, obesity, and high blood pressure; nor was it the result of less healthy men's shifting to sedentary jobs.

Another study followed the health status of approximately 17,000 male graduates of Harvard University for many years. All these men essentially had sedentary jobs, but they differed in the amount of leisure time they spent on physical activities. Men who expended at least 2,000 kilocalories per week on physical activity had only half the death rate from heart disease as did those men who expended less than 500 kilocalories per week. Not all of this energy was spent in exercise programs; some was expended during routine activities such as climbing stairs.

The effect of exercise on coronary-heart-disease risk factors. One of the important medical achievements of the 20th century has been the development of the risk-factor theory of coronary heart disease. Scientists have discovered that persons who are overweight, smoke cigarettes, have high blood pressure, or show elevated blood levels of certain types of fat- and cholesterol-carrying molecules are much likelier to die from coronary heart disease. Furthermore, combinations of these risk factors result in exponential increases in the risk of death. The discovery and description of risk factors have led to an understanding of the atherosclerotic process and of how to prevent and treat it. Evidence suggests that regular exercise can lower a person's exposure to several of the risk factors.

Fat and cholesterol are transported by the blood in complex molecules called lipoproteins. Researchers have identified several classes of lipoproteins and have elucidated their roles in atherosclerotic progression. It is, therefore, possible to describe abnormal, or high-risk, lipoprotein profiles. Diet and heredity are key factors determining a person's lipoprotein profile, and exercise also plays an important role. Regular aerobic exercise improves the lipoprotein profile in most individuals. Although more work is needed to completely understand this association, the dose of exercise necessary to effect a beneficial change in the lipoprotein profile seems to be about eight to 10 miles of running (or its equivalent in other activity) per week.

Elevated blood pressure (hypertension) is a second powerful risk factor for coronary heart disease. Sedentary living habits and low levels of physical fitness increase the risk of developing hypertension. Exercise also appears to lower blood pressure in at least some individuals with hypertension. The greatest benefit is probably for younger people (those less than 40 to 45 years of age) whose hypertension is of relatively recent onset.

Effect of exercise on high blood pressure

Excess body weight is considered by most experts to be an independent risk factor for coronary heart disease, although obesity also indirectly increases the risk via deleterious impact on blood pressure and the lipoprotein profile. Exercise habits are strongly related to body weight. In virtually all studies of large populations, the more active individuals weigh less. One of the most consistent results seen in exercise-training studies is the loss of body weight and fat. Weight-loss programs that incorporate exercise as well as diet are more successful than those that rely on diet alone.

Impact on other chronic diseases. Although more research is needed to arrive at definitive conclusions, some evidence has suggested that regular exercise may help in the treatment or prevention of other chronic diseases. The control of type II diabetes, for example, appears to be aided by regular exercise. This form of diabetes is a major health problem in which the patient shows elevated levels of blood sugar despite having acceptable levels of insulin, the hormone that normally clears the blood of excess sugar by facilitating its utilization by the body cells. Persons with this disease need to control their blood sugar, but not with insulin injections. Oral medications that lower blood-sugar levels are available, but their usefulness has been questioned. Consequently, dietary modifications and exercise, both of which can lower blood-sugar levels, have become the key measures in controlling type II diabetes. Exercise seems to improve the insulin sensitivity of cells, so that blood sugar can more readily be taken in and used as fuel.

Effect of exercise on diabetes

A few reports have linked low physical activity with a higher risk of developing certain cancers, particularly colon cancer. These results are intriguing, but more work is needed to firmly establish that sedentary habits are an independent risk factor for cancer.

RISKS OF EXERCISE

As can be seen from the foregoing discussion, regular participation in an exercise program can provide several benefits. Yet exercise is similar to other medical or health interventions in that there are also potential costs associated with the activity. These costs range from minor inconveniences, such as time taken up by exercise, to more serious complications, including injury and even sudden death.

Injuries. It is clear that some people who participate in exercise training will develop injuries to their bones, muscles, and joints. Despite unfounded reports in the mass media of extremely high injury rates among adult exercisers, there have been few good studies of exercise injuries in populations. One of the difficulties in performing such studies has been the need to identify both the number of cases (individuals who become injured) and the number of persons at risk for injury (the total number of individuals exercising in the population). These two figures are necessary in order to calculate true injury rates. The best available studies on injury rates suggest that about 25 to 30 percent of adult runners will become injured over the course of a year, if injury is defined as an incident that causes an individual to stop exercising for at least one week. If only more serious injuries, such as those for which the individual seeks medical care, are considered, injury rates are much lower, perhaps in the range of 1 percent per year.

Little is known about the causes of exercise injuries. One factor that has been linked to injury is the amount of exercise; for example, individuals who run more miles are likelier to be injured than those who run fewer miles. Factors such as age, sex, body type, and experience have not been shown to be associated with risk of injury. It seems logical that structural abnormalities, sudden increases in training intensity, and types of equipment used are likely to be related to injury risk; however, data to support these opinions are not available.

How to avoid injuries

In view of the limited scientific data on injury risk, the exerciser is advised to follow commonsense practices until such time as the causes of injury are better understood. Exercisers should start their program slowly and gradually progress to more intensive training levels. They should use good equipment and pay particular attention to proper footwear. Exercisers who have had previous injuries should recognize that they may be more susceptible to similar injuries in the future. All exercisers should use caution and should monitor their bodies for the early warning signs of injury. If a problem begins to develop, it is good advice to stop exercising or to reduce the intensity of training for a few days to see if the problem disappears. Exercisers should not be afraid to experiment on themselves to find out what training practices and techniques seem to be more comfortable and less likely to produce injury. Moderation is good advice: few injuries are reported in individuals who run 10 to 15 miles per week, and this level is adequate to provide many health benefits.

Sudden death. Obviously, the most serious complication from an exercise program is sudden death. This is, fortunately, a rare occurrence. As discussed earlier, several studies have shown that individuals who regularly participate in exercise have a lower risk of dying from a heart attack. There is, however, also evidence that suggests a higher risk of dying during exercise than during sedentary activities. When one considers the total risk of sudden death over a 24-hour period, regular exercisers are much less likely to experience this catastrophe.

Virtually all individuals who drop dead suddenly have advanced coronary heart disease. It follows, therefore, that the best way to reduce the risk of sudden death during exercise is to avoid getting advanced coronary heart disease. This implies following good health practices in other aspects of life, such as not smoking, eating a prudent diet, and maintaining an ideal body weight. Individuals who are middle-aged or older can probably reduce their risk of sudden death by knowing about their coronary risk status and their general state of health before undertaking an exercise program. There are, of course, no guarantees, but if an individual has a thorough examination by a competent physician, including a maximal exercise test and other procedures that screen for coronary heart disease, that person can probably safely begin an exercise program.

SUMMARY

There has been much progress in the field of exercise and physical conditioning. Concepts about exercise have moved from faddism to scientific legitimacy, thanks to researchers in physical education, exercise physiology, and medicine. Yet much remains to be learned, and experts need to work together to further develop the study and promotion of exercise. There are many items that need further study, from the cellular level to the population as a whole. For example, more information is needed on specifically how exercise affects blood lipoprotein levels, and further research is needed on rates of injuries in populations of exercisers.

BIBLIOGRAPHY. Exercise as a key to health maintenance is found in KENNETH H. COOPER, *The Aerobics Program for Total Well-being: Exercise, Diet, Emotional Balance* (1982); MICHAEL L. POLLOCK, JACK H. WILMORE, and SAMUEL M. FOX III, *Exercise in Health and Disease: Evaluation and Prescription for Prevention and Rehabilitation* (1984); PHILIP L. WHITE and THERESE MONDEIKA (eds.), *Diet and Exercise: Synergism in Health Maintenance* (1982); BUD GETCHELL and WAYNE ANDERSON, *Being Fit: A Personal Guide* (1982); JOHN E. BEAULIEU, *Stretching for All Sports* (1980). Specific aspects of exercise for middle-aged or older people are the topic of HERBERT A. DEVRIES and DIANNE HALES, *Fitness After 50* (1982). Other special topics are treated in the *Journal of the American Medical Association:* LARRY W. GIBBONS *et al.*, "The Acute Cardiac Risk of Strenuous Exercise," *J.A.M.A.*, 244(16):1799–1801 (Oct. 17, 1980); JOHN J. DUNCAN *et al.*, "The Effects of Aerobic Exercise on Plasma Catecholamines and Blood Pressure in Patients with Mild Essential Hypertension," *J.A.M.A.*, 254(18):2609–13 (Nov. 8, 1985); RALPH S. PAFFENBARGER *et al.*, "A Natural History of Athleticism and Cardiovascular Health," *J.A.M.A.*, 252(4):491–495 (July 27, 1984); and STEVEN N. BLAIR *et al.*, "Physical Fitness and Incidence of Hypertension in Healthy Normotensive Men and Women," *J.A.M.A.*, 252(4):487–490 (July 27, 1984). See also KENNETH H. COOPER, *Running Without Fear: How to Reduce the Risk of Heart Attack and Sudden Death During Aerobic Exercise* (1985); and SIDNEY ALEXANDER, *Running Healthy: A Guide to Cardiovascular Fitness* (1980).

What happens to the body during exercise and other intense physical activity is explained in many informative sources and texts. See PER-OLOF ASTRAND and KAARE RODAHL, *Textbook of Work Physiology: Physiological Bases of Exercise,* 2nd ed. (1977); and GEORGE A. BROOKS and THOMAS D. FAHEY, *Exercise Physiology: Human Bioenergetics and Its Applications* (1984). Public health aspects of physical activities and exercise are explored in a collection of articles in *Public Health Reports,* vol. 100, no. 2 (March–April 1985). ROY J. SHEPHARD (ed.), *Frontiers of Fitness* (1971), discusses the physiology of exercise and desirable limits of fitness for people of different ages.

The usefulness of recreational exercise was studied in Greek antiquity by Galen; see ROBERT MONTRAVILLE GREENE, *A Translation of Galen's "Hygiene" (De sanitate tuenda)* (1951). For a historical treatment of exercise and sport, see RICHARD D. MANDELL, *Sport, a Cultural History* (1984), a scholarly study of physical activity as a component of culture; WILLIAM J. BAKER, *Sports in the Western World* (1982); and HISPA (INTERNATIONAL ASSOCIATION FOR THE HISTORY OF PHYSICAL EDUCATION AND SPORT), *The History, the Evolution and Diffusion of Sports and Games in Different Cultures* (1976).

(K.H.C./S.N.B.)

Exploration

Of all the characteristics that human beings have used to distinguish themselves from "lower" animals, the desire to explore the unknown may be the most enduring. As the Norwegian polar explorer and oceanographer Fridtjof Nansen observed, "Man wants to know, and when he ceases to do so, he is no longer man."

The gradual expansion by certain animals into new ranges cannot be regarded as exploration. Rather, such movements are merely shifts to less crowded or otherwise more favourable environmental settings. To be sure, humans also have made such shifts when subjected to population and predatory pressures. Then, too, economic and military considerations often have been major driving factors in human expansion into new realms. Nonetheless, in numerous instances, human exploration has been marked by imaginative leaps across hostile stretches, sometimes at great risk, to reach something undefined simply for its own sake.

Much of the history of exploration—certainly of modern geographical exploration—has been European. Such has been the case not because Europeans possessed superior curiosity or some other internal force but because whatever events channelled their societies toward an advanced level of technology allowed them to expend more energy on exploration. The rapid growth and consolidation of their kingdoms provided them with an opportunity to exploit new discoveries fully. By contrast, the great Asian kingdoms, though no less capable, turned inward and erected walls—both legal and literal—between themselves and the "barbarians" of the outside world. The accomplishments of early European explorers are described in the article EUROPEAN OVERSEAS EXPLORATION AND EMPIRES, THE HISTORY OF.

The 20th century witnessed the last stages of exploration of the Earth's surface and the initial attempts to explore the deep sea and space. Scientific attention today is primarily directed toward these new frontiers. Efforts also are being made to investigate the interior of the Earth, knowledge of which still remains relatively limited. This article considers all three pursuits, highlighting the objectives, methods, and findings of each. (Da.D.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 723 and 738 and the *Index*.

The article is divided into the following sections:

Exploration of the Earth's surface and interior

By the beginning of the 20th century most of the Earth's surface had been explored, at least superficially, except for the Arctic and Antarctic regions.

Today the last of the unmarked areas on land maps have been filled in by radar and photographic mapping from aircraft and satellites. One of the last areas to be mapped was the Darién peninsula between the Panama Canal and Colombia. Heavy clouds, steady rain, and dense jungle vegetation made its exploration difficult, but airborne radar was able to penetrate the cloud cover to produce reliable, detailed maps of the area. In recent years data returned by Earth satellites have led to several notable discoveries, as, for example, drainage patterns in the Sahara, which are relics of a period when this region was not arid.

Historically, exploration of the Earth's interior was confined to the near surface, and this was largely a matter of following downward those discoveries made at the surface. Most present-day scientific knowledge of the subject has been obtained through geophysical research conducted since World War II, and the deep Earth remains a major frontier in the early 21st century.

Exploration of space and the ocean depths has been facilitated by the placement of sensors and related devices in these regions. Only a very limited portion of the subsurface regions of the Earth, however, can be studied in this way. Investigators can drill into only the uppermost crust, and the high cost of drilling is severely limiting. Because direct exploration is so restricted, investigators are forced to rely extensively on geophysical measurements (see below).

Problems of subsurface exploration

PRIMARY OBJECTIVES AND ACCOMPLISHMENTS

Scientific curiosity is a major motive for exploring the Earth's subsurface regions. Another key motive is the prospect of economic profit. Improved standards of living have increased the demand for water, fuel, and other materials, creating economic incentives. Pure knowledge has often been a by-product of profit-motivated exploration; by the same token, substantial economic benefits have resulted from the quest for scientific knowledge.

Many surface and subsurface exploratory projects are undertaken with the aim of locating: (1) oil and gas accumulations and coal beds; (2) concentrations of commercially important minerals (ores of iron, copper, and uranium) and deposits of building materials (sand, gravel, etc.); (3) recoverable groundwater; (4) various rock types at different depths for engineering planning; (5) geothermal reserves; and (6) archaeological features.

Concern for safety has prompted extensive searches for possible hazards before major construction projects are undertaken. Sites for dams, power plants, factories, tunnels, hazardous waste depositories, and so forth need to be stable and provide assurance that underlying formations will not shift or slide from the weight of the construction, move along a fault during an earthquake, or permit the seepage of water or wastes. Geophysical surveys furnish a more complete picture than test boreholes alone, although some boreholes are usually drilled to verify the geophysical interpretation.

METHODOLOGY AND INSTRUMENTATION

Geophysical techniques

Geophysical techniques involve measuring reflectivity, magnetism, gravity, acoustic or elastic waves, radioactivity, heat flow, electricity, and electromagnetism. Most measurements are made on the surface of the land or sea, but some are taken from aircraft or satellites, and still others are made underground in boreholes or mines and at ocean depths.

Geophysical mapping depends on the existence of a difference in physical properties of adjacent bodies of rock—*i.e.,* between whatever is being sought and those of the surroundings. Often the difference is provided by something associated with but other than whatever is being sought. Examples include a configuration of sedimentary layers that form a trap for oil accumulation, a drainage pattern that might affect groundwater flow, or a dike or host rock where minerals may be concentrated. Different methods depend on different physical properties. Which particular method is used is determined by what is being sought. In most cases, however, data from a combination of methods rather than from simply one method yield a much clearer picture.

Remote sensing. This comprises measurements of electromagnetic radiation from the ground, usually of reflected energy in various spectral ranges measured from aircraft or satellites. Remote sensing encompasses aerial photography and other kinds of measurements that are generally displayed in the form of photograph-like images. Its applications involve a broad range of studies, including cartographic, botanical, geological, and military investigations.

Fundamentals of remote sensing

Remote-sensing techniques involve using combinations of images. Images from different flight paths can be combined to allow an interpreter to perceive features in three dimensions, while those in different spectral bands may identify specific types of rock, soil, vegetation, and other entities, where species have distinctive reflectance values in different spectral regions (*i.e.,* tone signatures). Images taken at intervals make it possible to observe changes that occur over time, such as the seasonal growth of a crop or changes wrought by a storm or flood. Those taken at different times of the day or at different sun angles may reveal quite distinct features; for example, seafloor features in relatively shallow water in a calm sea can be mapped when the Sun is high. Radar radiation penetrates clouds and thus permits mapping from above them. Side-looking airborne radar (SLAR) is sensitive to changes in land slope and surface roughness. By registering images from adjacent flight paths, synthetic stereo pairs may give ground elevations.

Thermal infrared energy is detected by an optical-mechanical scanner. The detector is cooled by a liquid-nitrogen (or liquid-helium) jacket that encloses it, making the instrument sensitive at long wavelengths and isolating it from heat radiation from the immediate surroundings. A rotating mirror directs radiation coming from various directions onto the sensor. An image can be created by displaying the output in a form synchronized with the direction of the beam (as with a cathode-ray tube). Infrared radiation permits mapping surface temperatures to a precision of less than a degree and thus shows the effects of phenomena that produce temperature variations, such as groundwater movement.

Landsat images

Landsat images are among the most commonly used. They are produced with data obtained from a multispectral scanner carried aboard certain U.S. Landsat satellites orbiting the Earth at an altitude of about 900 kilometres (see also below *Applications satellites*). Images covering an area of 185 kilometres square are available for virtually every segment of the Earth's surface (see Plate 7). Scanner measurements are made in four spectral bands: green and red in the visible portion of the spectrum, and two infrared bands. The data are usually displayed by arbitrarily assigning different colours to the bands and then superimposing these to make "false-colour" images.

In geology, Landsat images are used to delineate landforms, rock outcrops and surface lithology, structural features, hydrothermal areas, and sites of mineral resources. Changes in vegetation revealed in the images may distinguish different soil types, subtle elevation differences, subsurface water distribution, subcropping rocks, and trace element distribution, among other things. Lineations of features may distinguish folded-rock strata or fault ruptures even where the primary features are not evident.

Magnetic methods. Measurements can be made of the Earth's total magnetic field or of components of the field in various directions. The oldest magnetic prospecting instrument is the magnetic compass, which measures the field direction. Other instruments include magnetic balances and fluxgate magnetometers. Most magnetic surveys are made with proton-precession or optical-pumping magnetometers, which are appreciably more accurate. The proton magnetometer measures a radio-frequency voltage induced in a coil by the reorientation (precession) of magnetically polarized protons in a container of ordinary water. The optical-pumping magnetometer makes use of the principles of nuclear resonance and cesium or rubidium vapour. It can detect minute magnetic fluctuations by measuring the effects of light-induced (optically pumped) transitions between atomic energy levels that are dependent on magnetic field strength.

Magnetic surveys from the air

Magnetic surveys are usually made with magnetometers borne by aircraft flying in parallel lines spaced two to four kilometres apart at an elevation of about 500 metres (one metre = 3.28 feet) when exploring for petroleum deposits and in lines 0.5 to one kilometre apart roughly 200 metres above the ground when searching for mineral concentrations. Ground surveys are conducted to follow up magnetic anomaly discoveries made from the air. Such surveys may involve stations spaced only 50 metres apart. Magnetometers also are towed by research vessels. In some cases, two or more magnetometers displaced a few metres from each other are used in a gradiometer arrangement; differences between their readings indicate the magnetic field gradient. A ground monitor is usually used to measure the natural fluctuations of the Earth's field over time so that corrections can be made. Surveying is generally suspended during periods of large magnetic fluctuation (magnetic storms).

Magnetics in petroleum and mineral prospecting

Magnetic effects result primarily from the magnetization induced in susceptible rocks by the Earth's magnetic field. Most sedimentary rocks have very low susceptibility and thus are nearly transparent to magnetism. Accordingly, in petroleum exploration magnetics are used negatively: magnetic anomalies indicate the absence of explorable sedimentary rocks. Magnetics are used for mapping features in igneous and metamorphic rocks, possibly faults, dikes, or other features that are associated with mineral concentrations. Data are usually displayed in the form of a contour map of the magnetic field, but interpretation is often made on profiles.

Rocks cannot retain magnetism when the temperature is above the Curie point (about 500° C for most magnetic materials), and this restricts magnetic rocks to the upper

40 kilometres of the Earth's interior. The source of the geomagnetic field must be deeper than this, and it is now believed that convection currents of conducting material in the outer core generate the field. These currents couple to the Earth's spin, so that the magnetic field—when averaged over time—is oriented along the planet's axis. The currents gradually change with time in a somewhat erratic manner and their aggregate effect sometimes reverses, which explains the time changes in the Earth's field. This is the crux of the magnetohydrodynamic theory of the geomagnetic field (see also EARTH: *Sources of the steady magnetic field*).

Gravimeters

Gravity methods. The gravity field of the Earth can be measured by timing the free fall of an object in a vacuum, by measuring the period of a pendulum, or in various other ways. Today almost all gravity surveying is done with gravimeters. Such an instrument typically consists of a weight attached to a spring that stretches or contracts corresponding to an increase or decrease in gravity. It is designed to measure differences in gravity accelerations rather than absolute magnitudes. Gravimeters used in geophysical surveys have an accuracy of about 0.01 milligal (mgal; 1 mgal = 0.001 centimetre per second per second). That is to say, they are capable of detecting differences in the Earth's gravitational ficld as small as one part in 100,000,000.

Gravity differences occur because of local density differences. Anomalies of exploration interest are often about 0.2 mgal. Data have to be corrected for variations due to elevation (one metre is equivalent to about 0.2 mgal), latitude (100 metres are equivalent to about 0.08 mgal), and other factors. Gravity surveys on land often involve meter readings every kilometre along traverse loops a few kilometres across. It takes only a few minutes to read a gravimeter, but determining location and elevation accurately requires much effort. Inertial navigation is sometimes used for determining elevation and location when helicopters are employed to transport gravimeters. Marine gravimeters are mounted on inertial platforms when used on surface vessels. A ship's speed and direction affcct gravimeter readings and limit survey accuracy. Aircraft undergo too many accelerations to permit gravity measurements except for regional studies.

In most cases, the density of sedimentary rocks incrcases with depth because the increased pressure results in a loss of porosity. Uplifts usually bring denser rocks nearer the surface and thereby create positive gravity anomalies. Faults that displace rocks of different densities also can cause gravity anomalies. Salt domes generally produce negative anomalies because salt is less dense than the surrounding rocks. Such folds, faults, and salt domcs trap oil, and so the detection of gravity anomalies associated with them is crucial in petroleum exploration. Moreover, gravity measurements are occasionally used to evaluate the amount of high-density mineral present in an ore body. They also provide a means of locating hidden caverns, old mine workings, and other subterranean cavities.

Seismic refraction methods. Seismic methods are based on measurements of the time interval between initiation of a seismic (elastic) wave and its arrival at detectors. The seismic wave may be generated by an explosion, a dropped weight, a mechanical vibrator, a bubble of high-pressure air injected into water, or other sources. The seismic wave is detected by a Geophone on land or by a hydrophone in water. An electromagnetic Geophone generates a voltage when a seismic wave produces relative motion of a wire coil in the field of a magnet, whereas a ceramic hydrophone generates a voltage when deformed by passage of a seismic wave. Data are usually recorded on magnetic tape for subsequent processing and display.

Seismic energy travels from source to detector by many paths. When near the source, the initial seismic energy generally travels by the shortest path, but as source-Geophone distances become greater, seismic waves travelling by longer paths through rocks of higher seismic velocity may arrive earlier (see Figure 1). Such waves are called head waves, and the refraction method involves their interpretation. From a plot of travel time as a function of source-Geophone distance, the number, thicknesses, and velocities of rock layers present can be determined for simple situations such as the one shown in Figure 2. The assumptions usually made are that (1) each layer is homogeneous and isotropic (*i.e.,* has the same velocity in all directions); (2) the boundaries (interfaces) between layers are nearly planar; and (3) each successive laycr has higher velocity than the one above. The velocity values determined from time-distance plots depend also on the dip (slope) of interfaccs, apparent velocities increasing whcn the Geophones are updip from the source and decreasing when downdip. By measuring in both directions the dip and rock velocity, each can be determined. With sufficicnt measurements, relief on the interfaces separating the layers also can be ascertained.

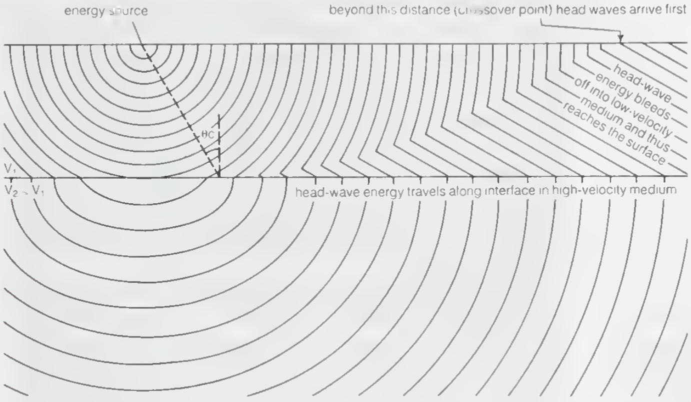

energy source
beyond this distance (crossover point) head waves arrive first
head-wave energy bleeds off into low-velocity medium and thus reaches the surface
θc
V_1
$V_2 > V_1$
head-wave energy travels along interface in high-velocity medium

Figure 1: Wave fronts (surfaces of disturbance at successive periods of time) showing how head-wave travel overtakes direct wave beyond some distance.

From R.E. Sheriff, *Encyclopedic Dictionary of Exploration Geophysics* (1984); published by the Society of Exploration Geophysicists

Fan shooting

High-velocity bodies of local extent can be located by fan shooting. Travel times are measured along different azimuths from a source, and an abnormally carly arrival time indicates that a high-velocity body was encountered at that azimuth. This method has been used to detcct salt domes, reefs, and intrusive bodies that are characterized by higher seismic velocity than the surrounding rock.

Two types of seismic waves can travel through a body: *P* waves (primary) and *S* waves (secondary). *P* waves are compressional waves and travel at the highest vclocity; hcnce, they arrive first. *S* waves are shcar waves that travel at a slower rate and are not able to pass through liquids that do not possess shear strength. In addition, there are

From R.E. Sheriff and L.P. Geldart, *Exploration Seismology. History, Theory, and Data Acquisition*, vol. 1 (1982); Cambridge University Press

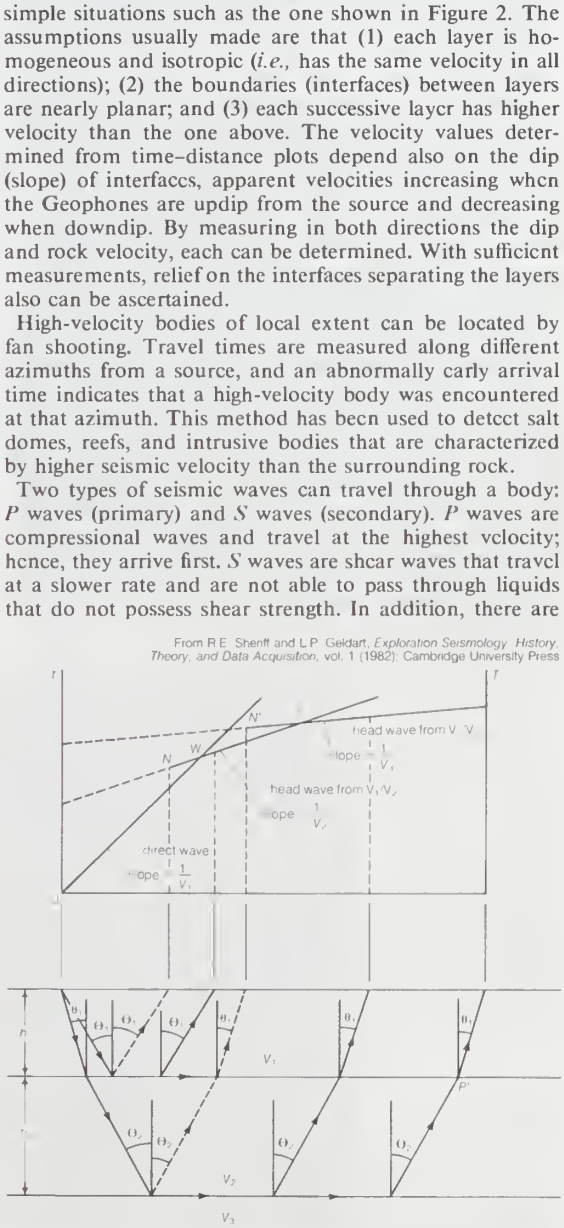


Figure 2: *Head waves from a three-layer situation.* (Top) Time–distance plot showing direct wave (*OW*) and head waves from the two interfaces (*NS* and *N'T*). (Bottom) Ray paths; $\Theta_1 = \sin^{-1}(V_1/V_2)$ = first critical angle, $\Theta_2 = \sin^{-1}(V_2/V_3)$ = second critical angle.

several types of seismic waves that can travel along surfaces. A major type of surface wave is the Rayleigh wave, in which a particle moves in an elliptical path in the vertical plane from the source. The horizontal component of Rayleigh waves is probably the principal cause of damage from earthquakes. Love waves are another type of surface wave; they involve shear motion. Still other varieties of surface waves can be transmitted through low-velocity layers (channel waves) or along the surface of a borehole (tube waves). Under certain circumstances (*e.g.*, oblique incidence on an interface), waves can change from one mode to another.

Significance of earthquake studies

Most of the current knowledge about the Earth's internal constitution is derived from analysis of the time–distance curves from earthquakes. Earthquakes usually generate several wave modes. These refract and reflect at interfaces within the Earth and partially change to other wave types to add to the number of seismic waves resulting from an earthquake. Different wave types can sometimes be distinguished by their components of motion detected by three-component seismographs; the direction from which they come can be determined by using an array of seismographs at the receiving station or by combining the data from different stations. The first wave motion from an earthquake reveals the nature of earth motion involved in the earthquake.

Very shallow seismic refraction is extensively used in engineering studies. Sometimes the energy source for shallow-penetration engineering studies involves simply hitting the ground with a sledgehammer. The ease with which a rock can be ripped by a bulldozer relates to the rock's seismic velocity. *S*-wave velocity measurements are of special interest to engineers because building stability depends on the shear strength of the foundation rock or soil. Seismic waves may be used for various other purposes. They are employed, for example, to detect faults that may disrupt a coal seam or fractures that may allow water penetration into a tunnel.

Seismic reflection methods. Most seismic work utilizes reflection techniques. Sources and Geophones are essentially the same as those used in refraction methods. The concept is similar to echo sounding: seismic waves are reflected at interfaces where rock properties change and the round-trip travel time, together with velocity information, gives the distance to the interface. The relief on the interface can be determined by mapping the reflection at many locations. For simple situations the velocity can be determined from the change in arrival time as source–Geophone distance changes.

In practice, the seismic reflection method is much more complicated. Reflections from most of the many interfaces within the Earth are very weak and so do not stand out against background noise. The reflections from closely spaced interfaces interfere with each other. Reflections from interfaces with different dips, seismic waves that bounce repeatedly between interfaces ("multiples"), converted waves, and waves travelling by other modes interfere with desired reflections. Also, velocity irregularities bend seismic rays in ways that are sometimes complicated.

The objective of most seismic work is to map geologic structure by determining the arrival time of reflectors (see Figure 3). Changes in the amplitude and waveshape, however, contain information about stratigraphic changes and occasionally hydrocarbon accumulations. In some cases, seismic patterns can be identified with depositional systems, unconformities, channels, and other features.

Advantage of seismic reflection

The seismic reflection method usually gives better resolution (*i.e.*, makes it possible to see smaller features) than other methods, with the exception of measurements made in close proximity, as with borehole logs (see below). Appreciably more funds are expended on seismic reflection work than on all other geophysical methods combined.

Electrical and electromagnetic methods. A multitude of electrical methods are used in mineral exploration. They depend on (1) electrochemical activity, (2) resistivity changes, or (3) permittivity effects. Some materials tend to become natural batteries that generate natural electric currents whose effects can be measured. The self-potential method relies on the oxidation of the upper surface of metallic sulfide minerals by downward-percolating groundwater to become a natural battery; current flows through the ore body and back through the surrounding groundwater, which acts as the electrolyte. Measuring the natural voltage differences (usually 50–400 millivolts [mV]) permits detecting continuous metallic sulfide bodies that lie astride the water table. Graphite, magnetite, anthracite, some pyritized rocks, and other phenomena also can generate self-potentials.

Self-potential method

The passage of an electric current across an interface where conduction changes from ionic to electronic results in a charge buildup at the interface. This charge builds up shortly after current flow begins, and it takes a short time to decay after the current circuit is broken. Such an effect is measured in induced-polarization methods and is used to detect sulfide ore bodies.

Resistivity methods involve passing a current from a generator or other electric power source between a pair of current electrodes and measuring potential differences

By courtesy of Seiscom Delta Inc

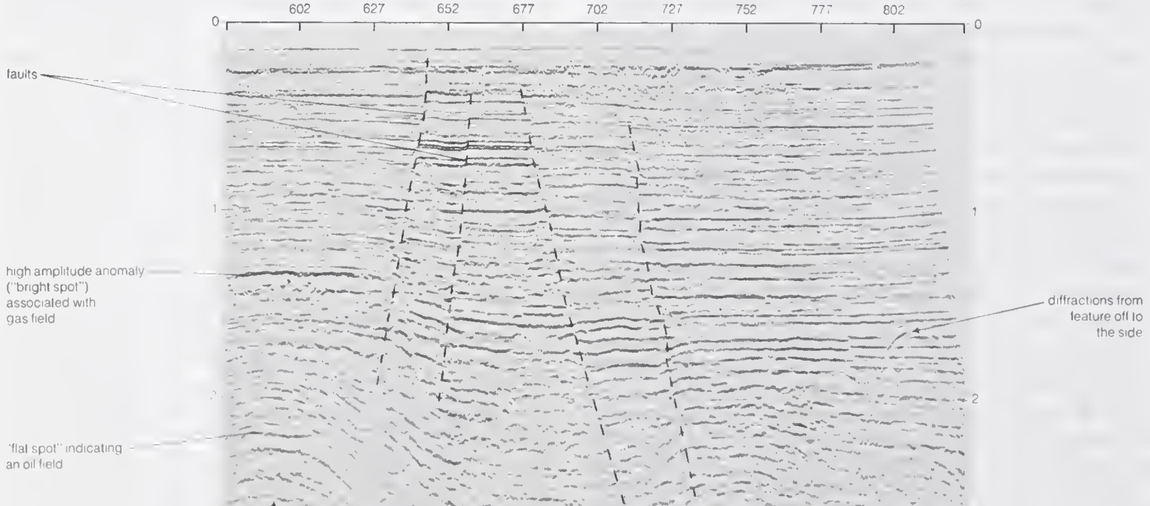


Figure 3: *Portion of a seismic time section from the Gulf of Mexico.*
The vertical scale is reflection arrival time. Section shows several faults indicated by disruptions in the reflections. At the right end of the section are a gas field at 1.350 seconds and an oil field at 2.200 seconds.

with another pair of electrodes. Various electrode configurations are used to determine the apparent resistivity from the voltage/current ratio. The resistivity of most rocks varies with porosity, the salinity of the interstitial fluid, and certain other factors. Rocks containing appreciable clay usually have low resistivity. The resistivity of rocks containing conducting minerals such as sulfide ores and graphitized or pyritized rocks depends on the connectivity of the minerals present. Resistivity methods also are used in engineering and groundwater surveys, because resistivity often changes markedly at soil/bedrock interfaces, at the water table, and at a fresh/saline water boundary.

Investigators can determine how resistivity varies over a given area by means of profiling methods, in which the location of an array of electrodes is altered but the same spacing between the component electrodes is maintained. Sounding methods enable investigators to pinpoint variations of resistivity with depth. In this case, electrode spacing is increased and, correspondingly, the effective depth of the contributing section. Several other techniques are commonly employed. Equipotential methods entail mapping equipotential lines that result from a current. Distortions from a systematic pattern indicate the presence of a body of different resistivity. The mise-a-la-masse method involves putting one current electrode in an ore body in order to map its shape and location.

The passage of current in the general frequency range of 500–5,000 hertz (Hz) induces in the Earth electromagnetic waves of long wavelength, which have considerable penetration into the Earth's interior. The effective penetration can be changed by altering the frequency. Eddy currents are induced where conductors are present, and these currents generate an alternating magnetic field, which induces in a receiving coil a secondary voltage that is out of phase with the primary voltage. Electromagnetic methods involve measuring this out-of-phase component or other effects, which makes it possible to locate low-resistivity ore bodies wherein the eddy currents are generated.

Magnetotelluric and related methods

Natural currents are induced in the Earth as a result of atmospheric disturbances (*e.g.,* lightning strikes) and bombardment of the upper atmosphere by the solar wind—a radial flow of protons, electrons, and nuclei of heavier elements emanating from the outer region of the Sun. Magnetotelluric methods measure orthogonal components of the electric and magnetic fields induced by these natural currents. Such measurements allow researchers to determine resistivity as a function of depth. The natural currents span a broad range of frequencies and thus a range of effective penetration depths. Related to the above techniques is the telluric-current method, in which the electric current variations are measured simultaneously at two stations. Comparison of the data permits determining differences in the apparent resistivity with depth at the two stations.

Electrical methods generally do not penetrate far into the Earth, and so do not yield much information about its deeper parts. They do, however, provide a valuable tool of exploring for many metal ores.

In addition, several electrical methods are used in boreholes. The self-potential (SP) log indicates mainly clay (shale) content, because an electrochemical cell is established at the shale boundary when the salinity of the borehole (drilling) fluid differs from that of the water in the rock. Resistivity measurements are made by using several electrode configurations and also by induction. Borehole methods are used to identify the rocks penetrated by a borehole and to determine their properties, especially their porosity and the nature of their interstitial fluids.

Radioactive methods. Radioactive surveys are used to detect ores or rock bodies associated with radioactive materials. Most natural radioactivity derives from uranium, thorium, and a radioisotope of potassium (potassium-40), as well as from radon gas. Radioactive elements are concentrated chiefly in the upper portion of the Earth's crust.

Radioactive disintegration, or decay, gives rise to spontaneous emission of alpha and beta particles and gamma rays. Detection is usually of gamma rays, and it is accomplished in most cases with a scintillometer, a photoconversion device containing a crystal of sodium iodide that emits a photon (minute packet of electromagnetic radiation) when struck by a gamma ray. The photon, whose intensity is proportional to the energy of the gamma ray, causes an adjacent photocathode to emit electrons, the exact number depending on the energy of the photon. The energy of the gamma ray itself is determined by the nature of the radioactive disintegration involved.

Where it can be assumed that a product element of a radioactive disintegration (a daughter isotope) is derived solely from the disintegration of a parent isotope that occurred after a rock's solidification (*i.e.,* as the rock cooled through its Curie point), the ratio of the parent/daughter isotopes present depends on the time since solidification. This often provides the basis for age determinations of rocks.

Gamma-ray logging

Information about the mineral composition and physical properties of a rock formation can be obtained by means of gamma-ray logging, a technique that involves measuring natural gamma-ray emissions in boreholes. In most sedimentary rocks, for example, potassium-40 is the principal emitter of gamma rays. Because potassium is generally associated with clays, a recording of gamma-ray emissions permits determination of clay (shale) content. In another related technique, the rock surrounding a borehole is bombarded by a radioactive source in the logging sonde and the effects of the reactions caused by the bombardment are measured. In a density log measurements are made of gamma rays that are backscattered from the rock formation, since their intensity indicates rock density. A neutron source is employed in another type of borehole log, one that is designed to reveal how much fluid occurs in a rock formation or how porous it is. Neutron energy loss is directly related to the density of protons (hydrogen nuclei) in rock, which is in turn reflective of its water content (or degree of porosity).

Geothermal methods. Temperature-gradient measurements are sometimes made to detect heat-flow anomalies; however, most exploration for geothermal resources (*e.g.,* superheated water and steam) is done with indirect methods. Resistivity or seismic methods, for example, may be used to map the magma chamber, which is the source of the heat, or to detect faults or other features that control the flow of hot subsurface water.

Measurement of trace elements

Geochemical methods. Since the early 1970s researchers have developed extremely sensitive methods of chemical analysis, providing the ability to detect minute amounts of materials. Many chemical elements are transported in very small quantities by fluids flowing in the Earth, so that a systematic measurement of such trace elements may help in locating their sources. Trace elements are sometimes associated with hydrocarbons (the principal constituents of petroleum, natural gas, and other fossil fuels); they can be utilized for identifying the specific types of hydrocarbons present in a given area.

Excavation, boring, and sampling. Direct sampling, usually by means of boreholes, is required to make positive identification of ores, fuels, and other materials. It is also necessary for determining their quantity and for selecting methods of recovery. Most deep boreholes are drilled by the rotary method, in which a drill bit is rotated while fluid ("drilling mud") is circulated through the bit to lubricate and cool it and to bring rock chips to the surface where they can be collected and analyzed. Shallow boreholes in hard rock formations are sometimes drilled by a percussion method, whereby a heavy bit is repeatedly raised and dropped to chip away pieces of rock. After a borehole has been drilled, various tools—sondes—are lowered into the hole to measure different physical properties.

CONCLUSIONS ABOUT THE DEEP EARTH

The overall oblate shape of the Earth was established by French Academy expeditions between 1735 and 1743. The Earth's mean density and total mass were determined by the English physicist and chemist Henry Cavendish in about 1797. It was later ascertained that the density of rocks on the Earth's surface is significantly less than the mean density, leading to the assumption that the density of the deeper parts of the planet must be much greater.

The Earth's magnetic field was first studied by William

Gilbert of England during the late 1500s. Since that time a long sequence of measurements has indicated its overall dipole nature, with ample evidence that it is more complex than the field of a simple dipole. Investigators also have demonstrated that the geomagnetic field changes over time. Moreover, they have found that magnetic constituents within rocks take on magnetic orientations as the rocks cool through their Curie point or, in the case of sedimentary rocks, as they are deposited. A rock tends to retain its magnetic orientation, so that measuring it provides information about the Earth's magnetic field at the time of the rock's formation and how the rock has moved since then. The field of study specifically concerned with this subject is called paleomagnetism.

From R E Sheriff, *Encyclopedic Dictionary of Exploration Geophysics* (1984), published by the Society of Exploration Geophysicists

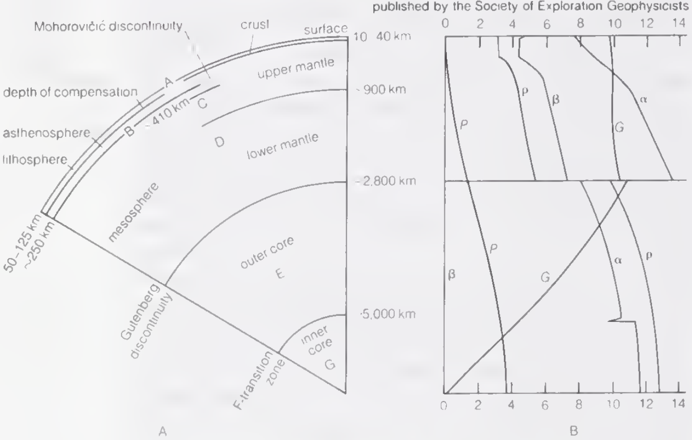


Figure 4: (A) Earth layering. (B) Variations of physical properties with depth in the Earth. P = pressure in 10^{11} Pa, ρ = density in g/cm³, β = S-wave velocity in km/s, α = P-wave velocity in km/s, G = gravity in Gal.

Observations of earthquake waves by the mid-1900s had led to a spherically symmetrical crust–mantle–core picture of the Earth (see Figure 4). The crust–mantle boundary is marked by a fairly large increase in velocity at the Mohorovičić discontinuity at depths on the order of 25–40 kilometres on the continents and five–eight kilometres on the seafloor. The mantle–core boundary is the Gutenberg discontinuity at a depth of about 2,800 kilometres. The outer core is thought to be liquid because shear waves do not pass through it.

Concept of plate tectonics

Scientific understanding of the Earth began undergoing a revolution from the 1950s. Theories of continental drift and seafloor spreading evolved into plate tectonics, the concept that the upper, primarily rigid part of the Earth, the lithosphere, is floating on a plastic asthenosphere and that the lithosphere is being moved by slow convection currents in the upper mantle. The plates spread from the mid-oceanic ridges where new oceanic crust is being formed, and they are destroyed by plunging back into the asthenosphere at subduction zones where they collide. Lithospheric plates also may slide past one another along strike-slip or transform faults (see also PLATE TECTONICS). Most earthquakes occur at the subduction zones or along strike-slip faults, but some minor ones occur in rift zones. The apparent fit of the bulge of eastern South America into the bight of Africa, magnetic stripes on the ocean floors, earthquake distribution, paleomagnetic data, and various other observations are now regarded as natural consequences of a single plate-tectonics model. The model has many applications; it explains much inferred Earth history and suggests where hydrocarbons and minerals are most likely to be found. Its acceptance has been widespread as economic conclusions have borne fruit.

An extensive series of boreholes drilled into the seafloor under the Joint Oceanographic Institutions for Deep Earth Sampling (JOIDES) program has established a relatively simple picture of the crust beneath the oceans (see also below *Undersea exploration*). In the rift zones where the plates comprising the Earth's thin crust separate, material from the mantle wells upward, cools, and solidifies. The molten mantle material that flows onto the seafloor and cools rapidly is called pillow basalt, while the underlying material that cools more slowly forms what is termed sheeted gabbro dikes. Sediments gradually accumulate on top of these, producing a comparatively simple pattern of sediment, basaltic basement, and gabbroic layering. Much of the heat flow from the solid Earth into the oceans results from the slow cooling of the oceanic basement rocks. Heat flow gradually declines with distance from the spreading centres (or with the length of time since solidification). As the oceanic rocks cool they become slightly denser, and isostatic adjustment causes them to subside slightly so that oceanic depths become greater. The oceanic crust is relatively thin, measuring only about five–eight kilometres in thickness. Nearly all oceanic rocks are fairly young, mostly Jurassic or younger (*i.e.*, less than 190,000,000 years old).

Seismic investigation of the continental crust

The crust beneath the continents, unlike the oceanic crust, is considerably older and thicker and appears to have been formed in a much more complex way. Because of its greater thickness, diversity, and complexity, the continental crust is much more difficult to explore. In 1975 the U.S. Geodynamics Committee initiated a research program to explore the continental crust using seismic techniques developed by private industry for the purpose of locating petroleum accumulations in sedimentary rocks. Since then its investigations have been conducted in a number of locales throughout the United States. Several notable findings have resulted from these studies, the most spectacular of which was the discovery of a succession of very low-angle thrust sheets beneath the Appalachian Mountains. This discovery, made from seismic data collected across this region between 1978 and 1980, is altering prevailing theories on continent formation.

The success of the U.S. crustal studies program has spawned a series of similar efforts in Australia, Europe, India, and elsewhere, and seismic investigation of the continental crust continues to be one of the most active areas of basic exploration in the last quarter of the 20th century.

The desire to detect nuclear explosions in the years following World War II led to the establishment of a worldwide network of uniform seismograph stations. This has greatly increased the number and reliability of earthquake measurements, the major source of information about the Earth's interior. The construction of large-array seismograph stations has made it possible to determine the directions of approach of earthquake waves and to sort out overlapping wave trains. Computer processing allows investigators to separate many wave effects from background noise and to analyze the implications of the multitude of observations now available.

The assumptions made in the past that significant property variations occur mainly in the vertical direction were clearly an oversimplification. Today, investigation of the deep Earth concentrates primarily on determining lateral (horizontal) changes and on interpreting their significance. Seismic tomographic analysis (see above) indicates appreciable variation in properties, many of which tend to confirm what is expected from theories about upper mantle convection currents. The data, however, also indicate variations whose significance has yet to be determined.

(R.E.Sh.)

Undersea exploration

PRIMARY OBJECTIVES AND ACCOMPLISHMENTS

The objectives of undersea exploration are to describe and understand the ocean waters, the seafloor, and the Earth beneath. Included in the scope of study are the physical and chemical properties of seawater, all manner of life in the sea, and the geological and geophysical features of the Earth's crust. Researchers in the field define and measure such properties; prepare maps in order to identify patterns; and utilize these maps, measurements, and theoretical models to achieve a better grasp of how the Earth works as a whole. This knowledge enables scientists to predict, for example, long-term weather and climatic changes and leads to more efficient exploration and exploitation of the Earth's resources, which in turn result in better management of the environment in general.

The multidisciplinary expedition of the British ship

First major undersea survey

"Challenger" in 1872–76 was the first major undersea survey. Although its main goal was to search for deep-sea life by means of net tows and dredging, the findings of its physical and chemical studies expanded scientific knowledge of temperature and salinity distribution of the open seas. Moreover, depth measurements by wire soundings were carried out all over the globe during the expedition.

Since the time of the "Challenger" voyage, scientists have learned much about the mechanics of the ocean, what it contains, and what lies below its surface. Investigators have produced global maps showing the distribution of surface winds as well as of heat and rainfall, which all work together to drive the ocean in its unceasing motion. They have discovered that storms at the surface can penetrate deep into the ocean and, in fact, cause deep-sea sediments to be rippled and moved. Recent studies also have revealed that storms called eddies occur within the ocean itself and that such a climatic anomaly as El Niño is caused by an interaction of the ocean and the atmosphere.

Other investigations have shown that the ocean absorbs large amounts of carbon dioxide and hence plays a major role in delaying its buildup in the atmosphere. Without the moderating effect of the ocean, the steadily increasing input of carbon dioxide into the atmosphere (due to the extensive burning of coal, oil, and natural gas) would result in the rapid onset of the so-called greenhouse effect—*i.e.*, a warming of the Earth caused by the absorption and reradiating of infrared energy to the terrestrial surface by carbon dioxide and water vapour in the air.

The field of marine biology has benefitted from the development of new sampling methods. Among these, broad ranging acoustical techniques have revealed diverse fish populations and their distribution, while direct, close up observation made possible by deep-sca submersibles has resulted in the discovery of unusual (and unexpected) species and phenomena.

In the area of geology, undersea exploration of the topography of the seafloor and its gravitational and magnetic properties has led to the recognition of global patterns of continental plate motion. These patterns form the basis of the concept of plate tectonics, which synthesized earlier hypotheses of continental drift and seafloor spreading. As noted earlier, this concept not only revolutionized scicntific understanding of the Earth's dynamic features (*e.g.*, seismic activity, mountain-building, and volcanism) but also yielded discoveries of economic and political impact. Earth scientists found that the mid-ocean centres of seafloor spreading also are sites of important metal deposits. The hydrothermal circulations associated with these centres produce sizable accumulations of metals important to the world economy, including zinc, copper, lead, silver, and gold. Rich deposits of manganese, cobalt, nickel, and other commercially valuable metals have been found in nodules distributed over the entirc ocean floor. The latter discovery proved to be a major factor in the establishment of the Convention of the Law of the Sea (1982), which calls for the sharing of these resources among developed and developing nations alike. Exploitation of these findings awaits only the introduction of commercially viable techniques for deep-sea mining and transportation.

Discovery of metal deposits on the seafloor

BASIC ELEMENTS OF UNDERSEA EXPLORATION

Platforms. Undersea exploration of any kind must be conducted from platforms, in most cases, ships, buoys, aircraft, or satellites. Typical oceanographic vessels capable of carrying out a full complement of underwater exploratory activities range in size from about 50 to 150 metres. They support scientific crews of 16 to 50 persons and generally permit a full spectrum of interdisciplinary studies. One example of a research vessel of this kind is the "Melville," operated by the Scripps Institution of Oceanography. It has a displacement of 2,075 tons and can carry 25 scientists in addition to 25 crew members. It is powered by a dual cycloidal propulsion system, which provides remarkable manoeuvrability.

Decp-sea drilling vessel

The "JOIDES Resolution," operated by Texas A & M University for the Joint Oceanographic Institutions for Deep Earth Sampling, represents a major advance in research vessels. A converted commercial drill ship, it measures 145 metres in length, has a displacement of 18,600 tons, and is equipped with a derrick that extends 62 metres above the waterline (see Figure 5). A computer-controlled dynamic positioning system enables the ship to remain over a specific location while drilling in water to depths as great as 8,300 metres. The drilling system of the ship is designed to collect cores from below the ocean floor; it can handle 9,200 metres of drill pipe. The vessel thus can sample most of the ocean floor, including the bottoms of deep ocean basins and trenches. The "JOIDES Resolution" has other notable capabilities. It can operate in waves as high as eight metres, winds up to 23 metres per second, and currents as strong as 1.3 metres per second. It has been outfitted for use in ice so that it can conduct drilling operations in high latitudes. The ship can accommodate 50 scientists as well as the crew and drilling team, and its geophysical laboratories total nearly 930 square metres.

Sub-mersibles

Other specialized vessels include the deep submergence research vehicle known as "Alvin" (see Figure 6), which can carry a pilot and two scientific observers to a depth of 4,000 metres. The manoeuvrability of the "Alvin" was pivotal to the discoveries of the mineral deposits at the mid-ocean seafloor spreading centres and of previously unknown biological communities living at those sites. Another versatile vessel is the Floating Instrument Platform (FLIP). It is a long narrow platform that is towed in a horizontal position to a research site. Once on location, the ballast tanks arc flooded to flip the ship to a vertical position. Only 17 metres of the ship extend above the waterline, with the remaining 92 metres completely submerged. The rise and fall of the waves cause a very small change in the displacement, resulting in a high degree of stability.

New ship designs that promise even greater stability and ease of use include that of thc Small Waterplane Area Twin Hull (SWATH) variety. This design type requircs the use of twin submerged, streamlined hulls to support a structure that rides above the watcr surface. The deck shape is entirely unconstrained by the hull shape, as is the case for conventional surface vessels. Ship motion is

By courtesy of Ocean Drilling Program, Texas A & M University

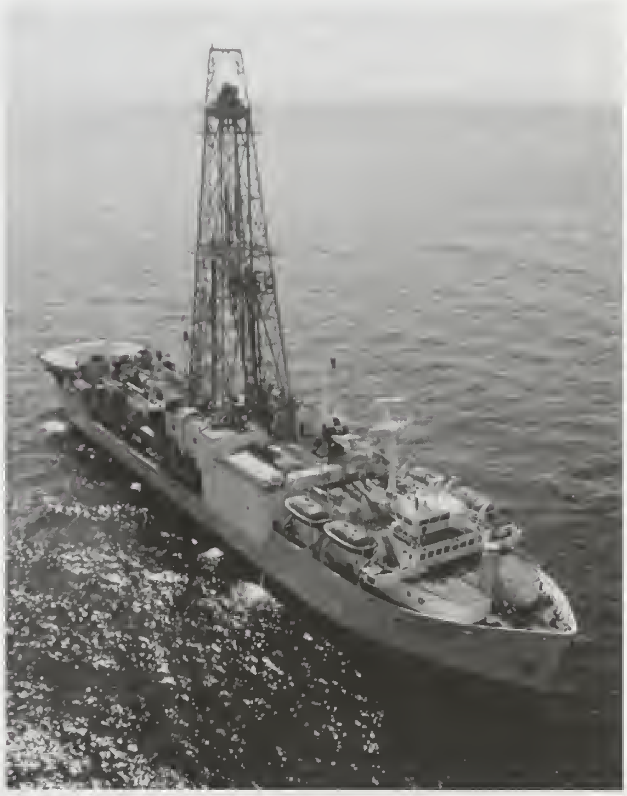


Figure 5: "JOIDES Resolution," a deep-sea drilling vessel that uses a computer-controlled, acoustic dynamic positioning system to maintain location over the drilling site. The derrick is visible amidships.

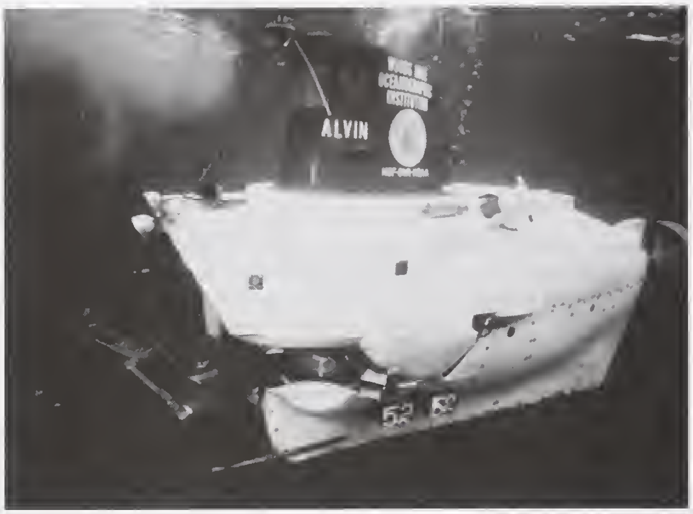


Figure 6: "Alvin," a small deep-diving submarine capable of submerging to depths of more than 4,000 metres.
By courtesy of Woods Hole Oceanographic Institution, photograph, Rod Catanach

greatly reduced because of the depth of the submerged hulls. For a given displacement, a SWATH-type vessel can provide twice the amount of deck space that a single-hull ship can, with only 10 percent of the motion of the single-hull design type. In addition, a large centre opening, or well, can be used to display and recover instruments.

Navigation. Exploration of any kind is useful only when the location of the discoveries can be noted precisely. Thus, navigation has always been a key to undersea exploration.

There are various ways by which the position of a vessel at sea can be determined. In cases where external references such as stars or radio and satellite beacons are unavailable or undetectable, inertial navigation, which relies on a stable gyroscope for determining position, is commonly employed. It is far more accurate than the long-used technique of dead reckoning, which is dependent on a knowledge of the ship's original position and the effects of the winds and ocean currents on the vessel.

Loran systems

Another modern position-fixing method is all-weather, long-range radio navigation. It was introduced during World War II as Loran (long-range navigation) A, a system that determines position by measuring the difference in the time of reception of synchronized pulses from widely spaced transmitting stations. The latest version of this system, Loran C, uses low-frequency transmissions and derives its high degree of accuracy from precise time-difference measurements of the pulsed signals and the inherent stability of signal propagation. Users of Loran C are able to identify a position with an accuracy of 0.4 kilometre and a repeatability of 15 metres at a distance of up to about 2,220 kilometres from the reference stations. The Loran C system covers heavily travelled regions in the North Pacific and North Atlantic oceans, parts of the Indian Ocean, and the Mediterranean Sea.

Satellite navigation

Satellite navigation has proved to be the most accurate method of locating geographical position. A polar-orbiting satellite system called Transit was established in the early 1960s by the United States to provide global coverage for ships at sea. In this system, a vessel pinpoints its position relative to a set of satellites whose orbits are known by measuring the Doppler shift of a received signal—*i.e.*, the change in the frequency of the received signal from that of the transmitted signal. The Transit system suffers from one major drawback. Because of the limited number of system satellites, the frequency with which position determinations can be made each day is relatively low, particularly in the tropics. The system is being improved to provide nearly continuous positioning capability at sea. This expanded version, the Global Positioning System (GPS), is to have 18 satellites, six in each of three orbital planes spaced 120° apart. The GPS is designed to provide fixes anywhere on Earth to an accuracy of 20 metres and a relative accuracy 10 times greater.

METHODOLOGY AND INSTRUMENTATION

Water sampling for temperature and salinity. The temperature, chemical environment, and movement and mixing of seawater are fundamental to understanding the physical, chemical, and biological features of the ocean and the geology of the ocean floor. Traditionally, oceanographers have collected seawater by means of specially adapted water-sampling bottles. The most universal water sampler used today, the Nansen bottle, is a modification of a type developed in the latter part of the 19th century by the Norwegian Arctic explorer and oceanographer Fridtjof Nansen. It is a metal sampler equipped with special closing valves that are actuated when the bottle, attached by one end to a wire that carries it to the desired depth, rotates about that end. A mercury thermometer fastened to the bottle records the temperature at the specified depth. The design of the device is such that, when it is inverted, its mercury column breaks. The amount of mercury remaining in the graduated capillary portion of the thermometer indicates the temperature at the point of inversion. This type of reversing thermometer and the Nansen bottle are extensively used by oceanographers because of their accuracy and dependability in a harsh environment.

The temperature and salinity of the ocean have been mapped with data gathered by many ships over many years. This information is used for tracing heat and water movement and mixing, as well as for making density measurements, which are employed in calculating ocean currents. It was noted as early as the "Challenger" expedition that the salt dissolved in seawater has remarkably constant major constituents. As a consequence, it is possible to map water density patterns within the sea with measurements of only the water temperature and one major property of the sea salt (*e.g.*, the chloride ion content or the electrical conductivity) to arrive at an accurate estimate of the density of a given sample.

The concept of standard seawater

Standard laboratory techniques such as titration are routinely used at sea for determining chlorinity. Chlorinity can be briefly defined as the number of grams of chlorine, bromine, and iodine contained in one kilogram of seawater, assuming that the bromine and iodine are replaced by chlorine. Salinity is the total weight of dissolved solids, in grams, found in one kilogram of seawater and may be determined from the concentration of chlorinity because of the constancy of major constituents. In the traditional technique, a solution of silver nitrate of a known strength is added to a sample of seawater to produce the same reaction as with "standard" seawater. The difference in the amounts added gives the degree of chlorinity. To ensure worldwide uniformity in chlorinity and salinity determinations, the International Council for the Exploration of the Sea prepared a universal reference, *Eau de Mer Normale* ("Standard Seawater"), in 1902. A new primary standard, prepared in 1937 and having a chlorinity of 19.381 parts per 1,000, is used to determine the chlorinities of all batches of standard seawater. It also is utilized to calibrate electrical conductivity measurements (see below).

Bathythermographs

Accurate and continuous measurements of temperature as it changes with depth are required for understanding how the ocean moves and mixes heat. To provide the necessary detail, temperature profilers had to be developed; then, with the introduction of reliable conductivity sensors, salinity profilers were added. An instrument called the bathythermograph (BT), which has been used since the early 1940s to obtain a graphic record of water temperature at various depths, can be lowered from a ship while it is moving at reduced speed. In this instrument a depth element (pressure-operated bellows) drives a slide of smoked glass or metal at right angles to a stylus. Actuated by a thermal element (liquid-filled bourdon tube) that expands and contracts in response to changes in temperature, the stylus scribes a continuous record of temperature and depth.

An expendable bathythermograph (XBT) was developed during the 1970s and has come into increasingly wider use. Unlike the BT, this instrument requires an electrical system aboard the research platform. It detects temperature variations by means of a thermistor (an electrical resistance element made of a semiconductor material) and depends on a known fall rate for depth determination. The sensor unit of the XBT is connected to the research platform by a leak-proof, insulated two-conductor cable. This cable is wound around a pair of large spools in an arrangement resembling that of a fisherman's spinning reel. In operation, the cable is unwound from each of the spools in a direction that is parallel to the axis of the respective spool. As a result, the cable unwinds from both the platform—either a ship or an airplane—and the sensor unit simultaneously but independently. Because of this double-spool arrangement, the sensor unit can free-fall from wherever it hits the sea surface and is completely unaffected by the direction or speed of the craft from which it was deployed. One of the principal reasons why the XBT has proved so useful is that it can provide a record of considerable depth even when it is deployed from a ship moving at full speed.

Electrical conductivity systems

Until the late 1950s, salinity was universally determined by titration. Since then, shipboard electrical conductivity systems have become widely used. Salinity-Temperature-Depth (STD) and the more recent Conductivity-Temperature-Depth (CTD) systems have greatly improved on-site hydrographic sampling methods. They have enabled oceanographers to learn much about small-scale temperature and salinity distributions.

The most recent version of the CTD systems features rapid-response conductivity and temperature sensors. The conductivity sensor consists of a tiny cell with four platinum electrodes. This type of conductivity cell virtually eliminates errors resulting from the polarization that occurs where the electrodes come in contact with seawater. The temperature sensor combines a tiny thermistor with a platinum-resistance thermometer. Its operations are carried out in such a way as to fully exploit the fast response of the thermistor and the high accuracy of the platinum thermometer. In addition, the system uses a strain gauge as a pressure sensor, the gauge being adjusted to reduce temperature effects to a minimum. This CTD system is extremely reliable. While its temperature precision is greater than 0.001° C over a range of −3° to +32° C, its conductivity precision is on the order of one part per million.

Electrical conductivity measurement of seawater salinity has been so effective that it has given rise to a new practical salinity scale, one that is defined on the basis of conductivity ratio. This scale has proved to be a more reliable way of determining density (*i.e.,* the weight of any given volume of seawater at a specified temperature) than the chlorinity scale traditionally used. Such is the case because chlorinity is ion specific while conductivity is sensitive to changes in any ion. Investigators have found that measurements of conductivity ratio make it possible to predict density with a precision almost one order of magnitude greater than was permitted by the chlorinity measurements of the past.

Studying dissolved gases in seawater

Water sampling for chemical constituents. Nutrient concentration (*e.g.,* phosphate, nitrate, silicate), the pH (acidity), and the proportion of dissolved gases are used by the ocean chemist to determine the age, origin, and movement of water masses and their effect on marine life. Analysis of dissolved gases, for example, is useful in tracing ocean mixing, in studying gas production in the ocean, and in elucidating the natural cycles of atmospheric pollutants. Many such measurements are conducted aboard ship by autoanalyzers, devices that continually monitor a flow of seawater by spectral techniques. Those analyses that cannot be accomplished by an autoanalyzer are carried out with discrete samples in shipboard or shore-based laboratories.

Radioactive chemical tracers are of special interest. Radioisotopes serve as time clocks, thus offering a means of determining the age of water masses, the absolute rates of oceanic mixing, and the generation and destruction of plant tissue. The distribution of these time clocks is controlled by the interaction of physical and biological processes, and so these influences must be disentangled before the clocks can be read. A notable example is the use of carbon-14 (^{14}C). Today, a number of oceanographic laboratories make carbon-14 measurements of oceanic dissolved carbon for the study of mixing and transport processes in the deep ocean. Until recently large samples of water—about 200 litres (one litre = 0.264 gallon)—were required for analysis. New techniques use a linear accelerator (a device that greatly increases the velocity of electrically charged atomic and subatomic particles) as a sophisticated mass spectrometer to directly determine abundancy ratios of carbon-14/carbon-13/carbon-12 atoms. The advantage of the newer methodology is that only very small sample amounts—about 250 millilitres (one millilitre = 0.034 fluid ounce), are required for high accuracy measurements.

Measurements of ocean currents. Ocean currents can be measured indirectly through data on density and directly with current meters. In the indirect technique, water density is computed from temperature and salinity observations, and pressure is then calculated from density. The resulting highs and lows of ocean pressure can be used to estimate ocean currents. The indirect technique establishes currents relative to a particular pressure surface; it is best for large-scale, low-frequency currents.

Direct measurement of currents is used to establish absolute currents and to monitor rapidly varying changes. In order to measure currents directly, a current meter must accurately record the speed and direction of flow, and the platform or mooring has to be reliable, readily deployable, and extremely sturdy. Researchers are able to make continuous measurements of currents at levels below the surface layer for periods of more than a year.

A typical system for the direct measure of ocean currents (see Figure 7) has three principal components: a surface or near-surface float; a line consisting of segments of wire and nylon that holds the current meters; and a release mechanism, signalled acoustically, which will drop an anchor when the system is ready to be brought back. A current meter typically employs a rotor equipped with a small direction vane that moves freely in line with the meter.

Impact of technological advances

One of the most important advances in modern instrument design has been the introduction of low-power, solid-state microelectronics. The accuracy of the Vector Averaging Current Meter (VACM), for example, has been

By courtesy of Woods Hole Oceanographic Institution

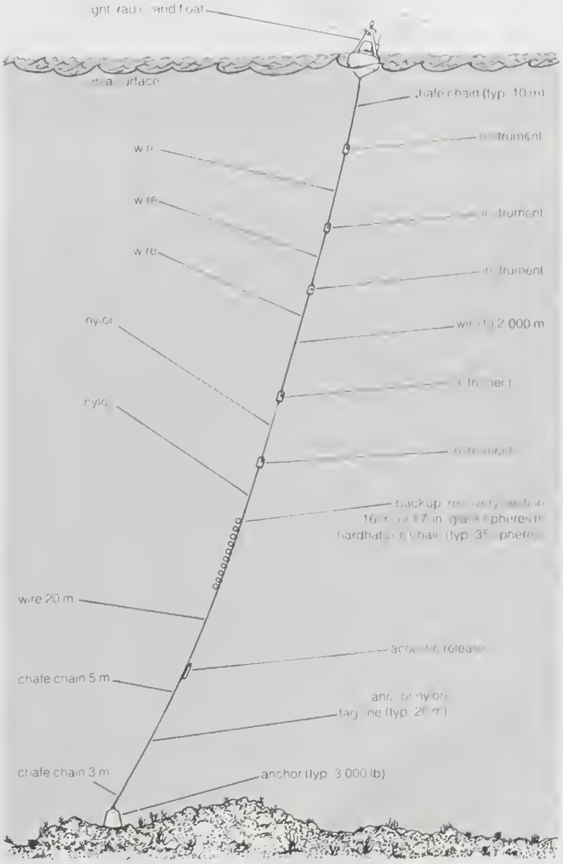


Figure 7: The surface mooring configuration used for measuring ocean currents in the deep sea by the Woods Hole Oceanographic Institution Buoy group.

improved appreciably by the use of integrated circuits, as has its data-handling capability. Because of the latter, the VACM can sample the direction and speed of currents roughly eight times during each revolution of the rotor. It then computes the north and east components of speed and stores this data, together with direction and time measurements, on a compact cassette recorder. The VACM is capable of making accurate measurements in wave fields as well as from moorings at the ocean surface because of its direct vector-averaging feature.

Currents also can be measured by drifting floats, either at the surface or at a given depth. Tracking the location of the floats is critical. Surface floats can be followed by satellite, but subsurface drifters must be tracked acoustically. A drifter of this sort acts as an acoustical source and transmits signals that can be followed by a ship with hydrophones suspended into the sea. For such tracking, a low sound frequency is crucial because the higher the frequency of sound, the more rapidly is its energy absorbed by the sea. The longest range floats available during the mid-1980s operated at a frequency as low as about 250 hertz. Long-range floats usually drift along channels known as sound fixing and ranging (SOFAR) channels, which occur in various areas of the ocean where a particular combination of temperature and pressure conditions affect the speed of sound. In a sense, the SOFAR channel acts as a type of acoustic waveguide that focusses sound; as a consequence, several watts of sound can be detected as far away as 2,000 kilometres or so.

Measuring vertical velocity in the ocean posed a major problem for years because of the difficulty of devising a platform that does not move vertically. During the 1960s oceanographers finally came up with a solution: they employed a neutrally buoyant float for measuring vertical velocities. This form of vertical-current meter consists of a cylindrical float on which fins are mounted at an angle. When water moves past the float, it causes the float to turn on its axis. Measurement of the rotation in relation to a compass yields the amount of vertical water movement.

Vertical-current meter

An extension of the neutrally buoyant float is the self-propelled, guided float. One such system, called a Self-Propelled Underwater Research Vehicle (SPURV), manoeuvres below the surface of the sea in response to acoustic signals from the research vessel. It can be used to produce horizontal as well as vertical profiles of various physical properties.

A Doppler-sonar system for measuring upper-ocean current velocity transmits a narrow beam that scatters off drifting plankton and other organisms in the uppermost strata of the ocean. From the Doppler shift of the backscattered sound, the component of water velocity parallel to the beam can be determined to a range of 1,400 metres from the transmitter with a precision of one centimetre per 0.1 second (one centimetre = 0.394 inch).

Integral to a complete picture of the ocean is a profile of velocity. Various methods have been devised for measuring currents as dependent on and varying with depth or horizontal position. Three techniques have been developed to make such measurements. The first involves acoustically tracking a "sinking float" as it descends toward the seafloor. The second technique entails the use of a free-fall device equipped with a current sensor. The third involves a class of current meter specially designed to move up and down a fixed line attached to a vessel, mooring, or drifting buoy. One such instrument has a roller block that couples the front of the instrument to a wire from the vessel. In this way, the motion of the vessel is decoupled from that of the instrument. Another important component of this instrument is its hull, a structure that not only furnishes buoyancy but also serves as a direction vane. In the bottom of the hull is a device that records velocity, temperature, and depth. The entire system descends at a rate of approximately 10 centimetres per second, resulting in a vertical resolution of several metres for the velocity profile produced.

Acoustic and satellite sensing. Remote sensing of the ocean can be done by aircraft and Earth-orbiting satellites or by sending acoustic signals through it. These techniques all offer a more sweeping view of the ocean than can be provided by slow-moving ships and hence have become increasingly important in oceanographic research.

Satellite-borne radar altimeters have proved to be especially useful. A radar system of this type can determine the distance between the satellite and the sea surface to an accuracy of better than 10 centimetres by measuring the time it takes for a transmitted pulse of radio energy to travel to the surface and return. By combining such a precise distance measurement with information about the satellite's orbit, oceanographers are able to produce maps of sea-surface topography. Moreover, they can deduce the pressure field of the sea surface by combining the distance measurement with knowledge about the geoid. They can in turn extrapolate information about the general circulation of the upper stratum of the ocean from a synoptic view of the surface pressure field.

Use of radar altimeters

Another remote-sensing technique involves the use of satellite-borne infrared and microwave radiometers to measure radiant energy released from the surface of the ocean. Such measurements are used to determine sea-surface temperature. High-resolution, infrared images transmitted by polar-orbiting satellites have provided researchers with an effective means of monitoring wave features in ocean currents over a wide area, as, for example, long equatorial waves in the Pacific Ocean and time variations in the flow of the Gulf Stream between Florida and Cape Hatteras, North Carolina.

Acoustic techniques also have many applications in the study of the ocean, particularly of those subsurface processes and physical properties inaccessible to satellite observation. In one such technique, the temperature structure of a water column from a given point on the seafloor to the surface is studied using an inverted echo sounder. This instrument, which features both an acoustic transmitter and a receiver, measures the time taken by a pulse of sound to travel from the sea bottom to the surface and back again. In most cases, a change in the average temperature of the water column above the instrument causes a fluctuation in the time interval between the transmission and the reception of the acoustic signal.

Other acoustic techniques can be utilized to study ocean variables on a large scale. A method known as ocean acoustic tomography, for example, monitors the travel time of sound pulses with an array of echo-sounding systems. In general, the amount of data collected is directly proportional to the product of the number of transmitters and receivers, so that much information on averaged oceanic properties can be gathered within a short period of time at relatively low cost.

Ocean acoustic tomography

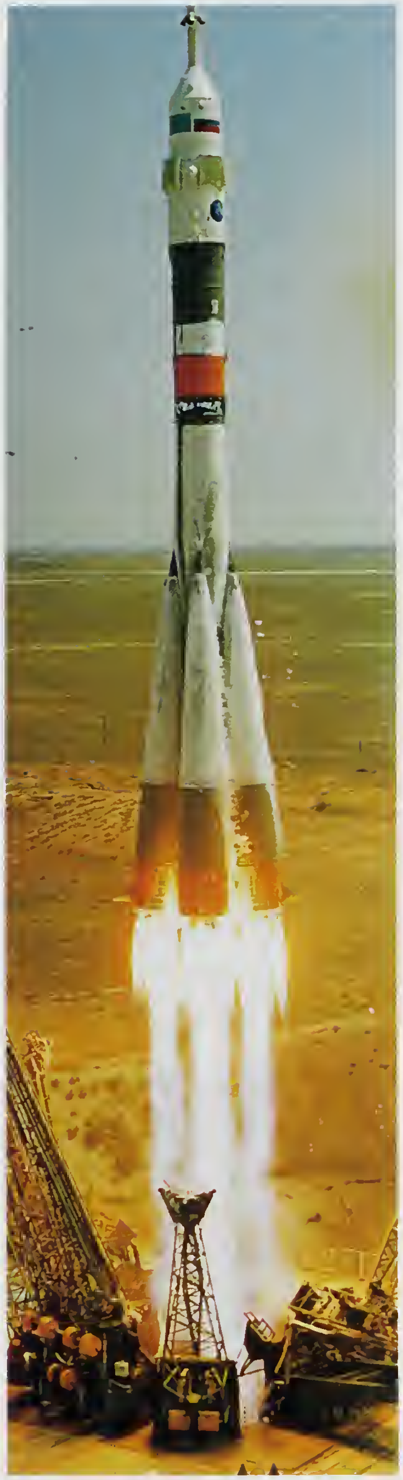
Collection of biological samples. Life at the bottom, benthos, is affected by the water column and by the sediment–water interface; the swimmers, or nekton, are influenced by the water that they come in contact with; and the floaters, or plankton—phytoplankton (plant forms) and zooplankton (animal forms)—are influenced by the water and the transfers that occur at the surface of the sea. Thus, in most cases, measurements and sampling of marine life is best done in concert with measurements of the physical and chemical properties of the ocean and the surface effects of the atmosphere.

As a consequence of the close interaction of sea life and its environment, marine biologists and biological oceanographers use most of the techniques mentioned above as well as some specialized techniques for biological sampling. Investigative techniques include the use of sampling devices, remote sensing of surface life-forms by satellite and aircraft, and in situ observation of plants and animals in direct interaction with their environment. The latter is becoming increasingly important as biologists recognize the fragility of organisms and the difficulty of obtaining representative samples. The absence of good sampling techniques means that even today little is known about the distribution, number, and life cycles of many of the important species of marine life.

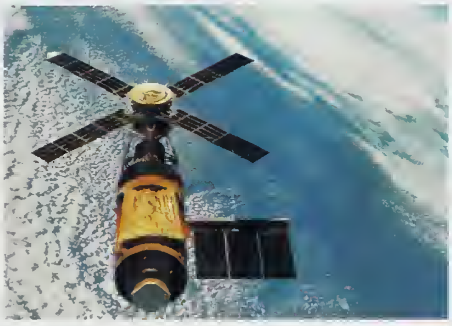
Some of the most commonly used samplers are plankton nets and midwater trawls. Nets have a mesh size smaller than the plankton under investigation; trawls filter out only the larger forms. The smaller net sizes can be used only when the ship is either stopped or moving ahead

Plankton nets and midwater trawls

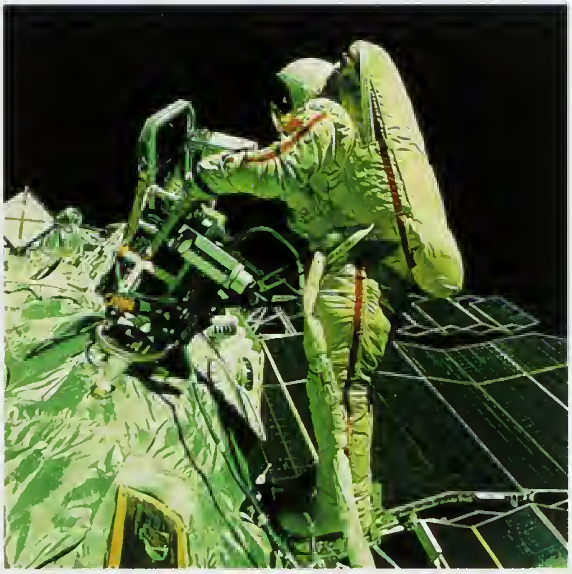
Space stations



Russian Soyuz TM-15 spacecraft atop its Soyuz launch vehicle during liftoff from the Baikonur Cosmodrome in Kazakhstan, July 22, 1992. On board is a Russian-French crew bound for the Mir space station.

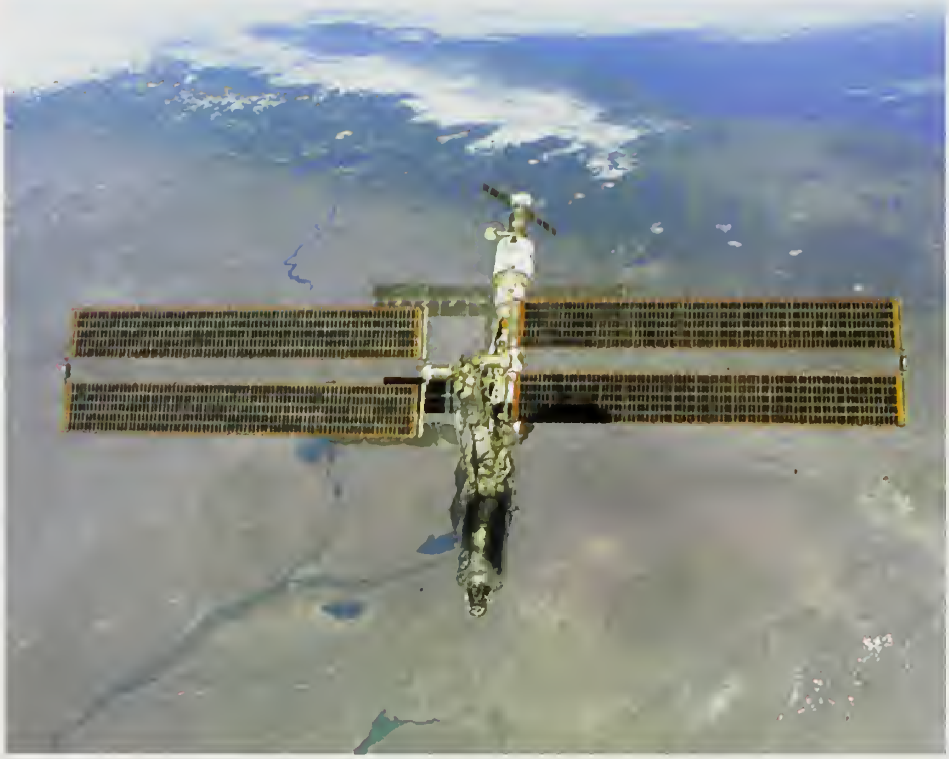


U.S. Skylab in Earth orbit, photographed February 8, 1974, by the final astronaut crew from the Skylab 4 Command Module. The makeshift sun shield and underlying "parasol" on the main part of the space station were installed by the first two crews to cover damage done to Skylab's protective shielding during launch.



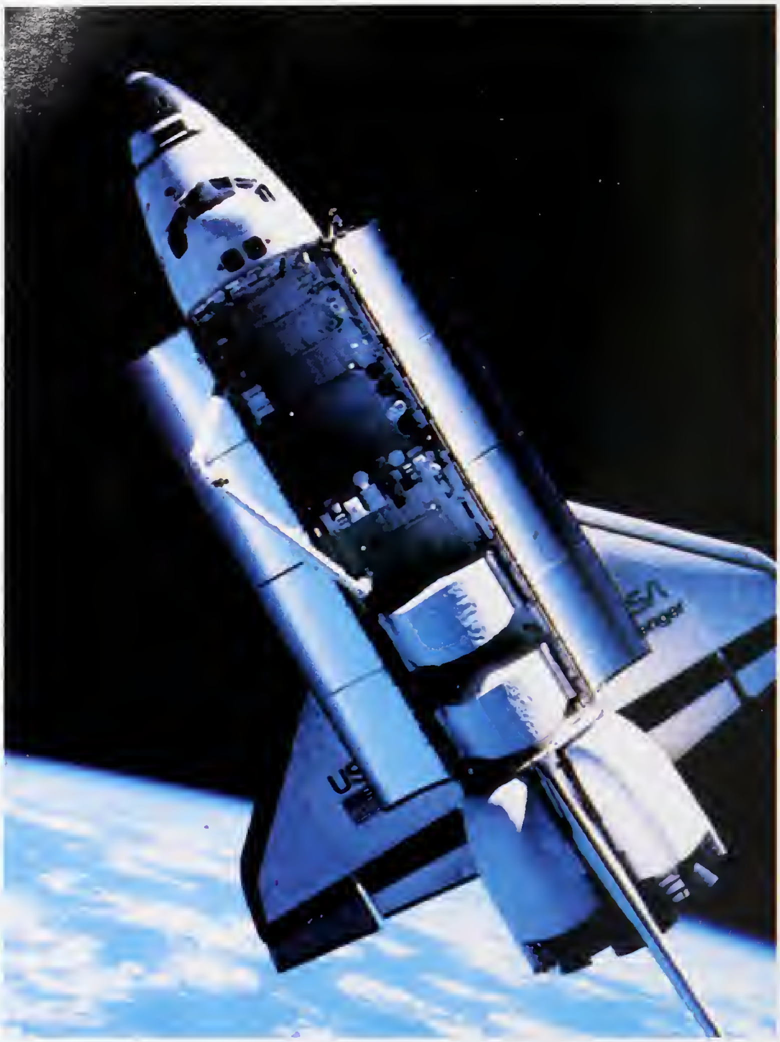
Soviet cosmonaut Svetlana Savitskaya making exterior repairs to the Salyut 7 space station, July 25, 1984. During her mission to Salyut 7, Savitskaya became the first woman to complete a space walk.

Plate 1 (Top left and middle right) Sovfoto/Eastfoto, (top right and bottom) by courtesy of National Aeronautics and Space Administration



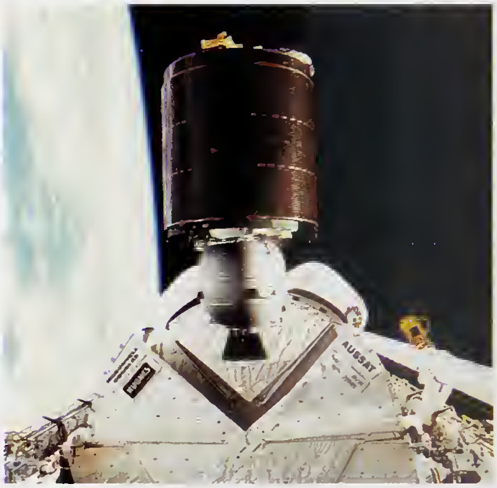
The International Space Station (ISS) photographed against the Negro River, Argentina, from the shuttle orbiter *Atlantis*, February 16, 2001. *Atlantis*'s primary mission was to deliver the Destiny laboratory module, visible at the leading (lower) end of the station.

Space shuttle

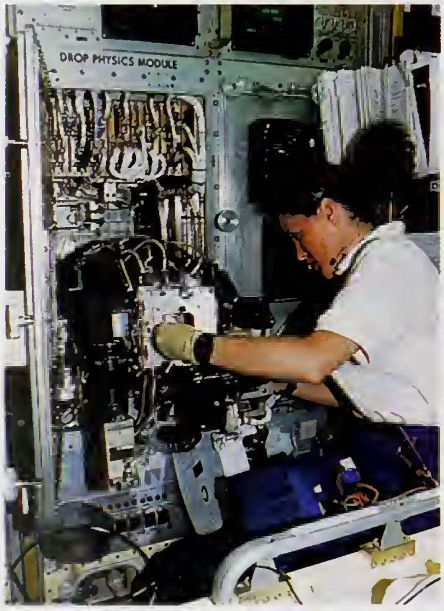


Open cargo bay of the shuttle orbiter *Challenger*, photographed February 7, 1984, by astronaut Bruce McCandless during a test of the manned maneuvering unit, a jet-equipped backpack that allows spacewalking astronauts to make powered excursions from their craft.

Liftoff of the U.S. space shuttle on its first mission, April 12, 1981, from the John F. Kennedy Space Center, Cape Canaveral, Florida. The shuttle orbiter is *Columbia*.



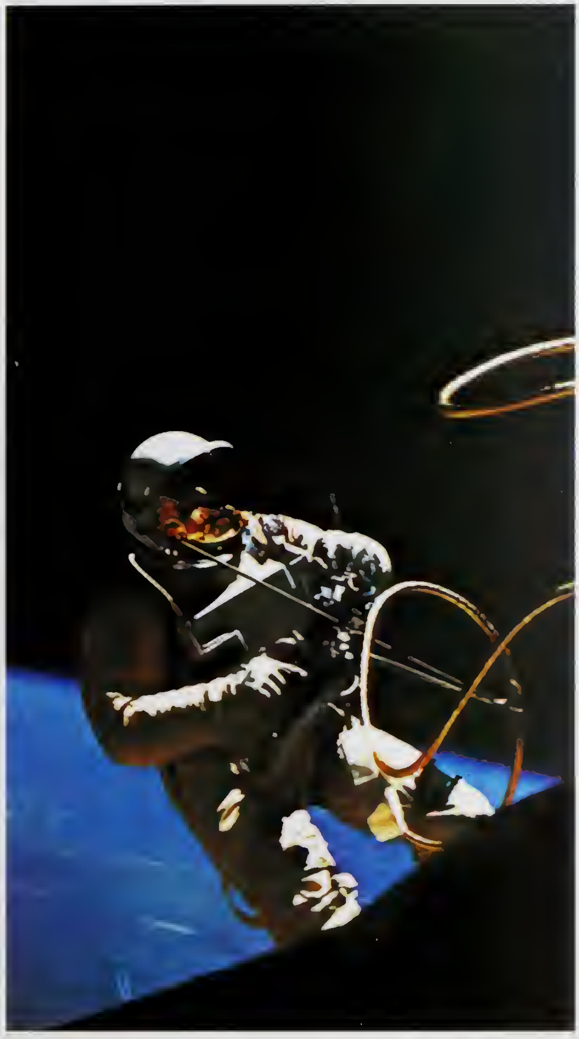
Australia's AUSSAT-1 communications satellite being deployed from the payload bay of the shuttle orbiter *Discovery*, August 27, 1985. The satellite subsequently was boosted into a geostationary orbit by means of an attached rocket motor.



U.S. astronaut Kathryn Thornton conducts an experiment in fluid physics aboard the shuttle orbiter *Columbia* during the U.S. Microgravity Laboratory mission flown October 20–November 5, 1995.

[illegible] National Aeronautics and Space Administration

Beginnings of human spaceflight



U.S. Gemini 4 astronaut Edward White during his historic 20-minute space walk on June 3, 1965. White was secured to the Gemini spacecraft by an umbilical and tether line and used a maneuvering gun to facilitate movement in space.

Soviet launch vehicle used to place manned Vostok spacecraft into orbit. Based on the R-7 intercontinental ballistic missile, the launcher had four strap-on liquid-fuel boosters surrounding the liquid-fuel core rocket.

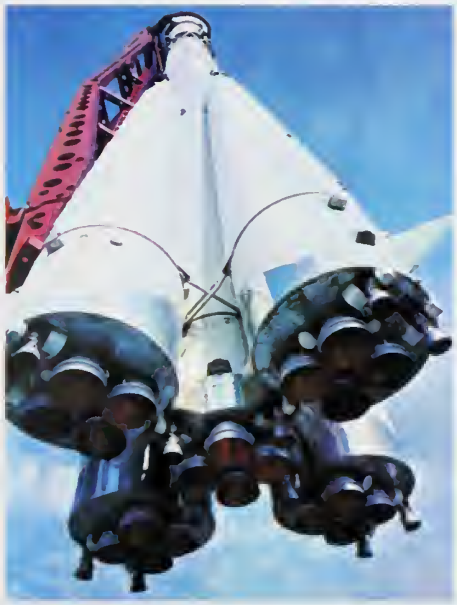
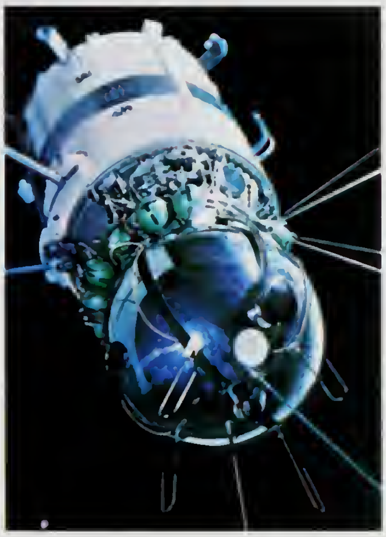
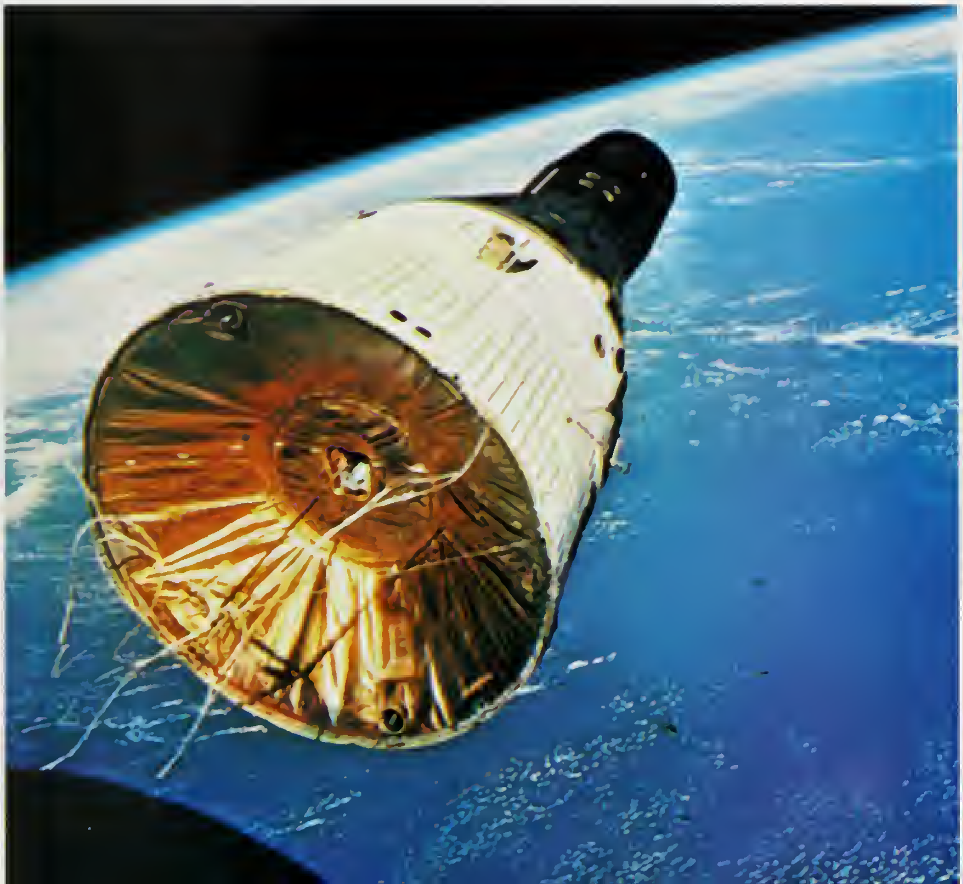


Plate 3 (Top left and bottom right) By courtesy of National Aeronautics and Space Administration, (top right and bottom left) Novosti Press Agency

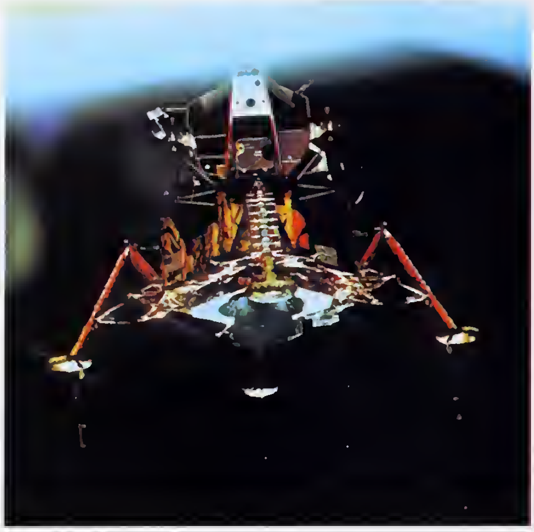


Soviet Vostok 6 spacecraft in which the first woman in space, cosmonaut Valentina Tereshkova, orbited Earth for three days. Vostok 6 was launched on June 16, 1963.

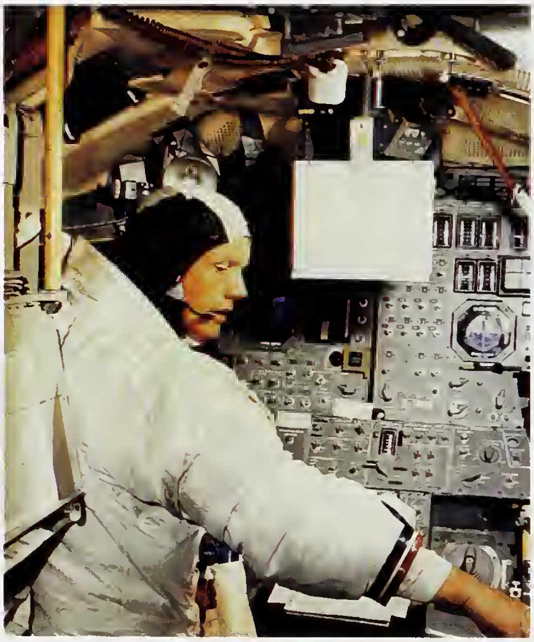
U.S. Gemini 7 spacecraft as photographed from Gemini 6, during rendezvous maneuvers. Gemini 7 was launched on December 4, 1965 and Gemini 6 on December 15



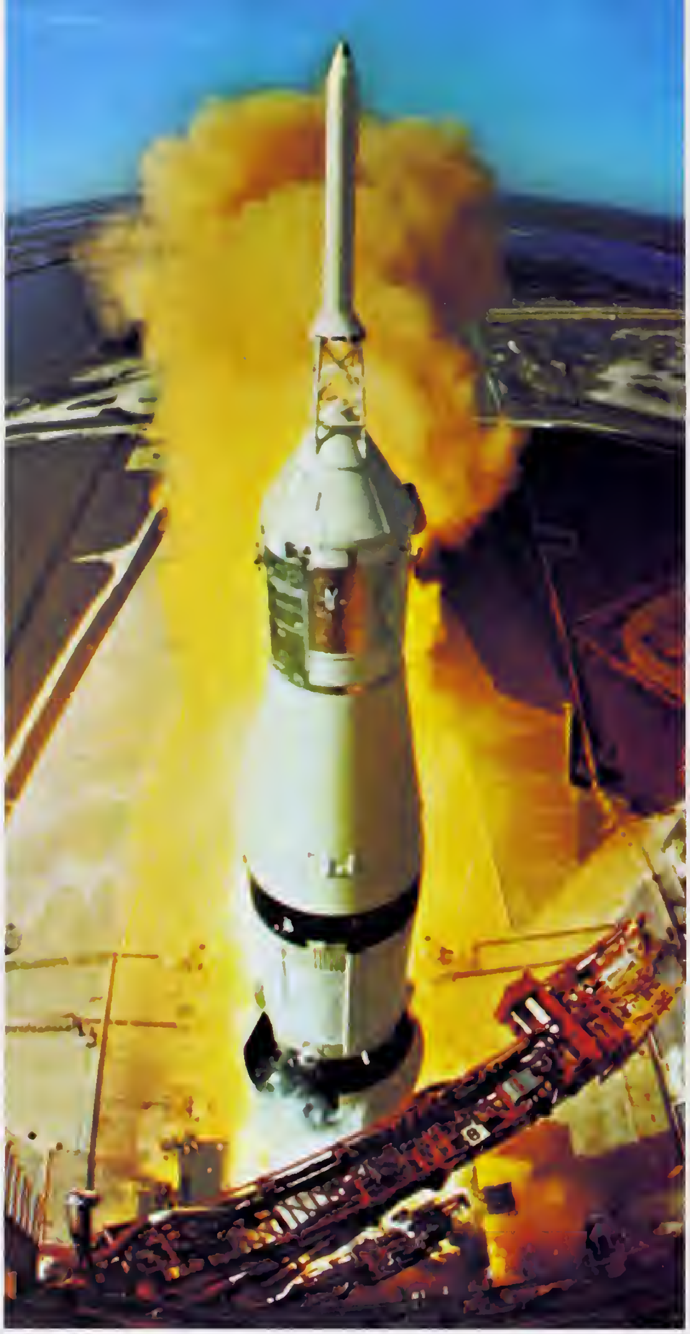
Apollo program



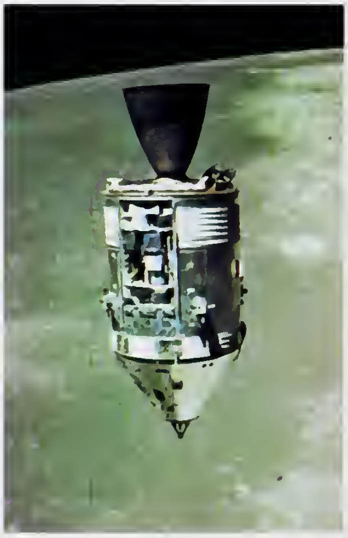
Apollo 11 Lunar Module with its four landing-gear footpads deployed for touchdown. This photograph was taken from the Command Module as the two spacecraft moved apart above the Moon on July 20, 1969.



Astronaut Neil Armstrong practicing in the Apollo Lunar Module Mission Simulator at the John F. Kennedy Space Center, Cape Canaveral, Florida, June 19, 1969, in preparation for the Apollo 11 lunar landing.



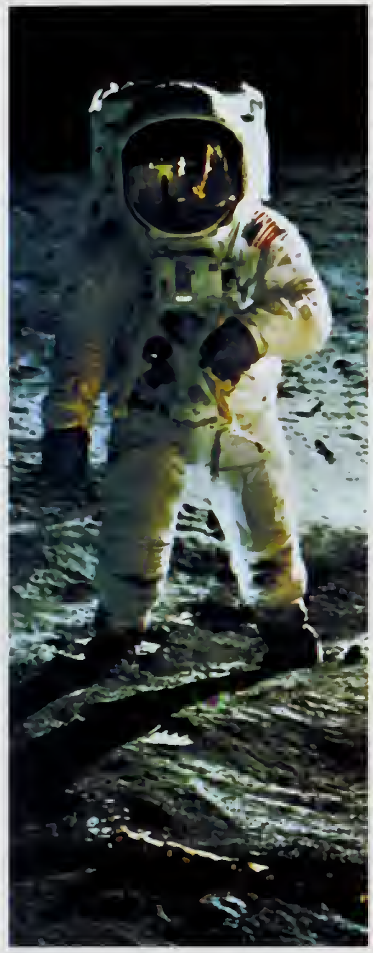
Apollo 15 spacecraft atop its Saturn V three-stage launch vehicle during liftoff from Cape Canaveral, Florida, on July 26, 1971. A camera mounted at the mobile launch tower's 110-metre (360-foot) level recorded the photograph.



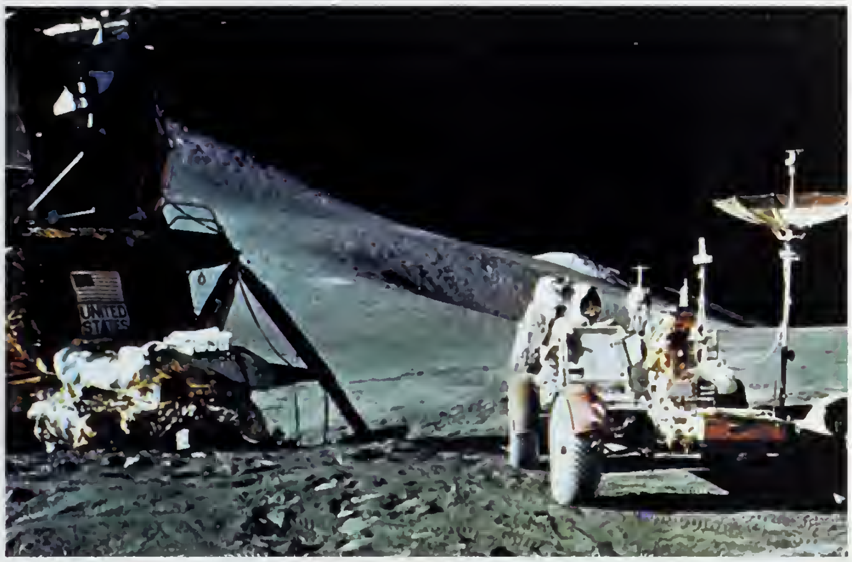
Apollo 15 Command and Service modules in lunar orbit with the Moon's surface in the background, as photographed from the Lunar Module.

Plate 4 By courtesy of National Aeronautics and Space Administration

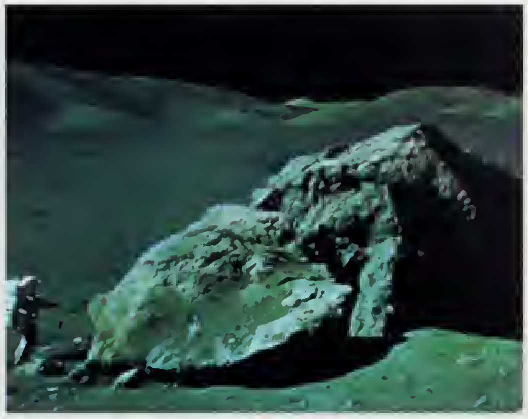
Humans on the Moon



Apollo 11 astronaut Edwin Aldrin, photographed July 20, 1969, during the first visit by human beings to the surface of the Moon. Reflected in Aldrin's helmet faceplate is the Lunar Module and astronaut Neil Armstrong, who took the picture.

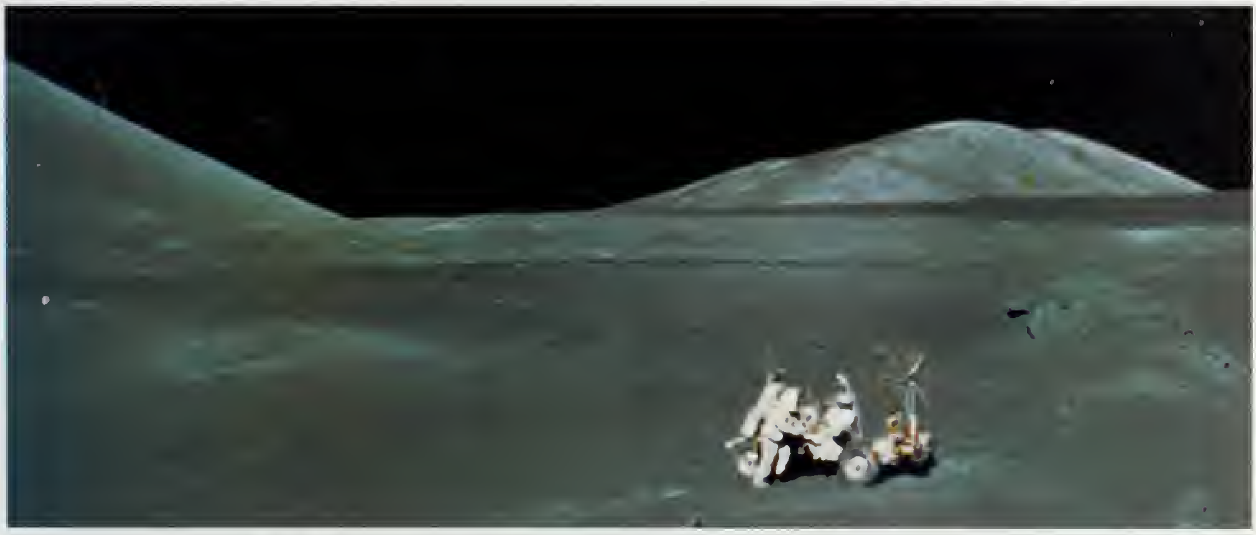


Apollo 15 astronaut James Irwin standing behind the Lunar Roving Vehicle, with the Lunar Module at left, on July 31, 1971, at the Hadley-Apennine landing site. St. George crater lies about five kilometres (three miles) behind Irwin's head.



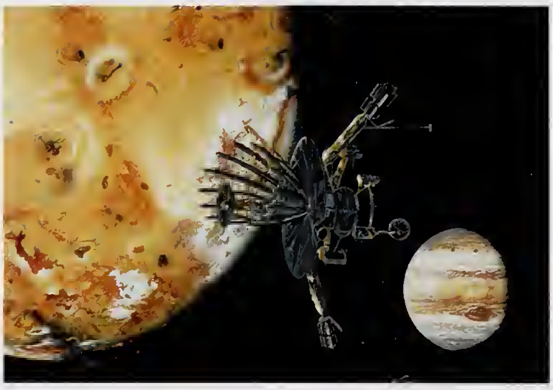
Apollo 17 geologist-astronaut Harrison Schmitt at the foot of a huge split boulder, December 13, 1972, during the mission's third extravehicular exploration of the Taurus-Littrow Valley landing site.

Plate 5 By courtesy of National Aeronautics and Space Administration



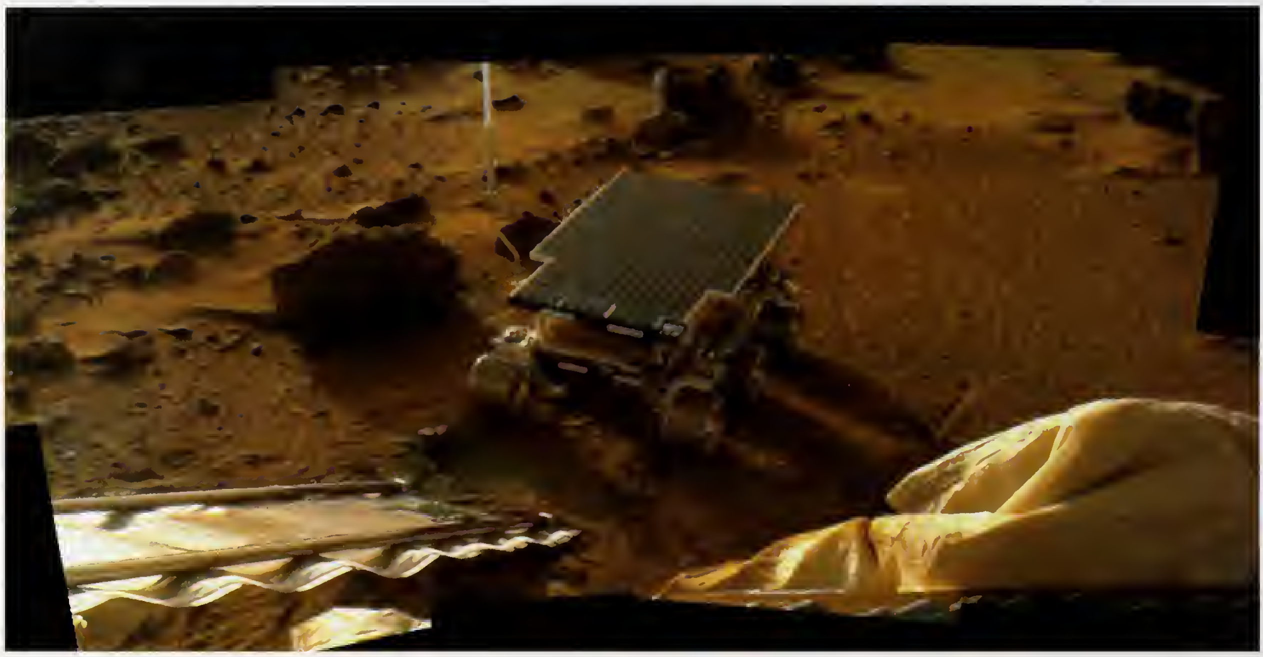
Apollo 17 geologist-astronaut Harrison Schmitt with the Lunar Roving Vehicle in the Moon's desolate Taurus-Littrow Valley, December 12, 1972. Harrison and fellow astronaut Eugene Cernan used the lunar rover in three extravehicular excursions, totaling 36 kilometres (22 miles), around the landing site of their Lunar Module.

Lunar and planetary spacecraft

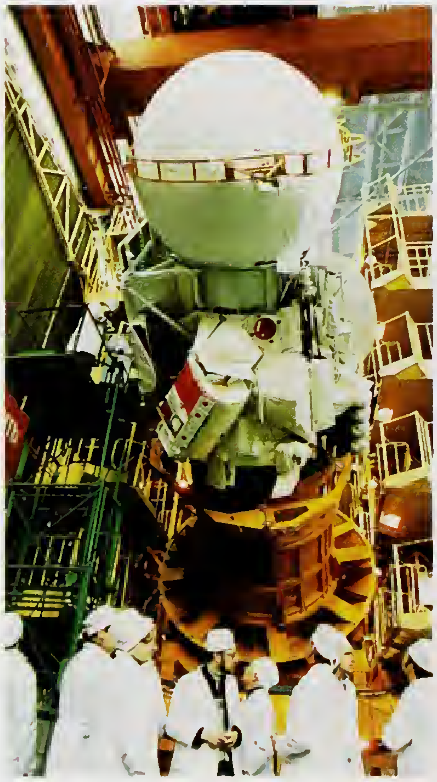


U.S. Galileo spacecraft, shown in an artist's conception, making a flyby of Jupiter's moon Io, with Jupiter in the background. Galileo, launched by space shuttle on October 18, 1989, went into orbit around Jupiter in December 1995.

U.S. Sojourner robotic roving vehicle, photographed by the Mars lander Pathfinder after deployment on the Martian surface, July 6, 1997. The Pathfinder/Sojourner spacecraft, launched on December 4, 1996, returned thousands of images of Mars and carried out a variety of studies.



Japanese Nozomi spacecraft approaching Mars, in an artist's conception. Launched on July 4, 1998, from the Kagoshima Space Center, Nozomi was designed to study the Martian upper atmosphere from orbit over a two-year period.

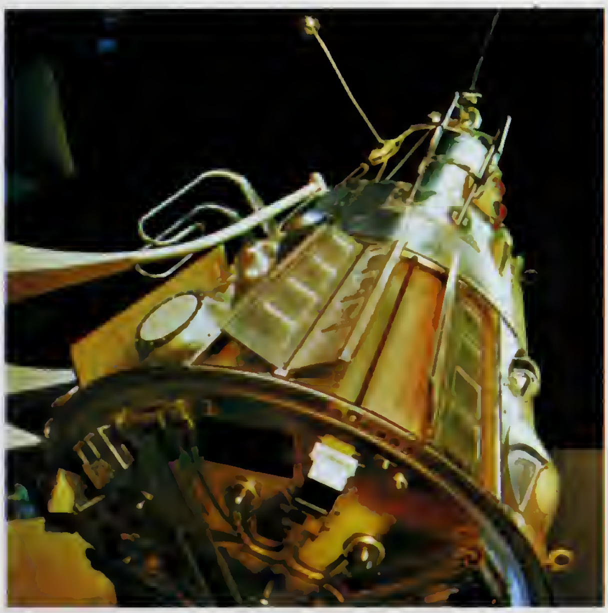


Soviet Vega spacecraft under assembly at the Baikonur Cosmodrome in Kazakhstan. Vega 1 and 2 were launched to Venus on December 15 and 21, 1984, respectively. After dropping balloon-borne lander probes into the Venusian atmosphere, they traveled on for flybys of Halley's Comet in 1986.

Plate 6 (Top) Illustration by Michael W Carroll, (middle) Jet Propulsion Laboratory/National Aeronautics and Space Administration, (bottom left) © The Institute of Space and Astronautical Science, (bottom right) Sovfoto/Eastfoto

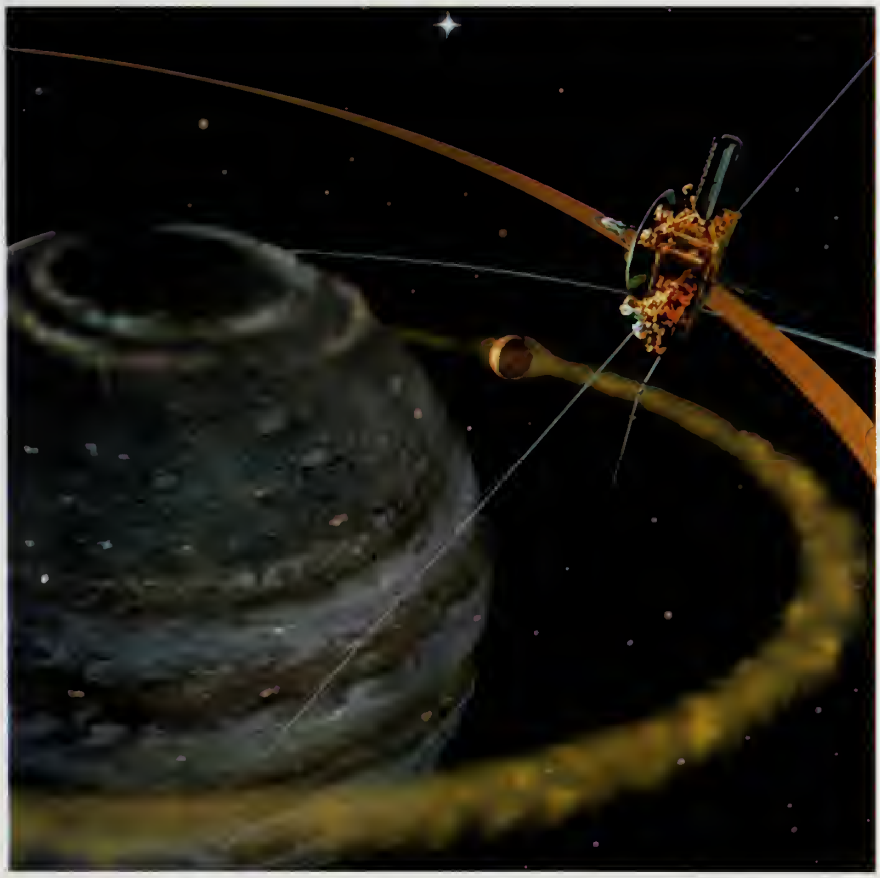
Scientific satellites and probes

Soviet Sputnik 3, the first multipurpose space-science satellite to be placed in orbit. Developed for the International Geophysical Year of 1957–58 and launched on May 15, 1958, it made and transmitted measurements of Earth's upper atmosphere, charged particles, and primary cosmic rays.



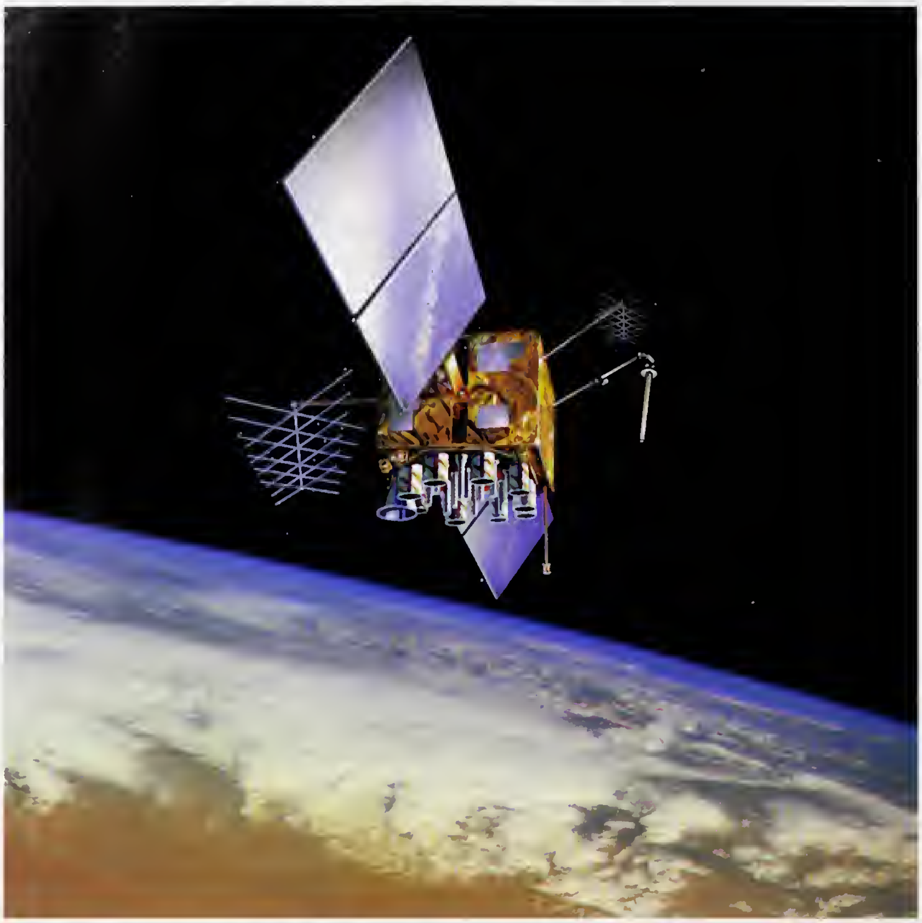
Netherlands-U.S.-U.K. Infrared Astronomical Satellite (IRAS), depicted in a cutaway model. Launched on January 25, 1983, the Earth-orbiting observatory mapped the sky for 10 months at infrared wavelengths above the interference of Earth's atmosphere.

(Above) U.S. Near Earth Asteroid Rendezvous (NEAR) spacecraft in orbit around an asteroid, in an artist's conception. Launched on February 17, 1996, NEAR rendezvoused with the asteroid Eros, which it studied for a year in orbit before touching down on its surface in February 2001.



European Space Agency's Ulysses spacecraft, in an artist's conception, shown engaged in a gravity-assist, or slingshot, maneuver near Jupiter in February 1992 that placed it in an orbit passing over the poles of the Sun. Ulysses was launched by space shuttle on October 6, 1990.

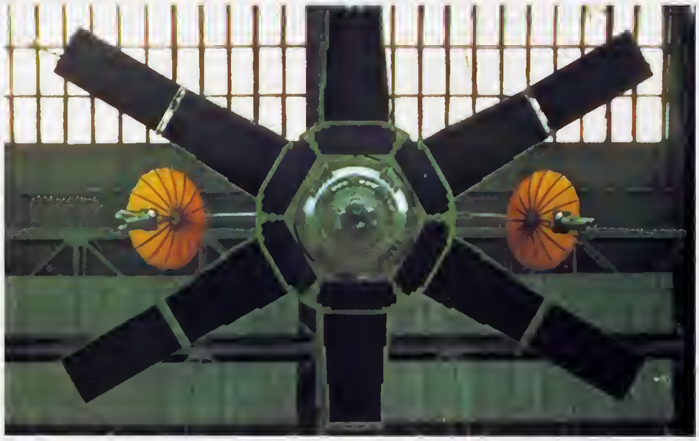
Plate 7 (Top left) Tass/Sovfoto (top right) Roger Ressmeyer—Corbis, (middle right) The Johns Hopkins University Applied Physics Laboratory/Jet Propulsion Laboratory/National Aeronautics and Space Administration (bottom) illustration by David A. Hardy



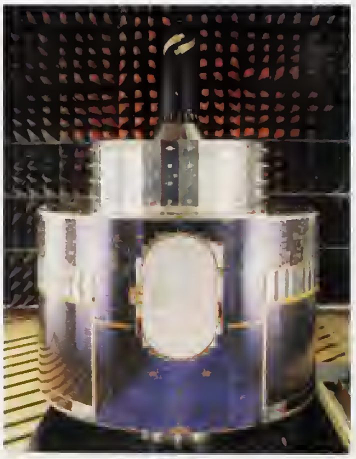
U.S. Navstar Global Positioning System (GPS) satellite, shown in an artist's conception. Deployment of the full complement of 24 GPS satellites was completed in 1994.

Applications satellites

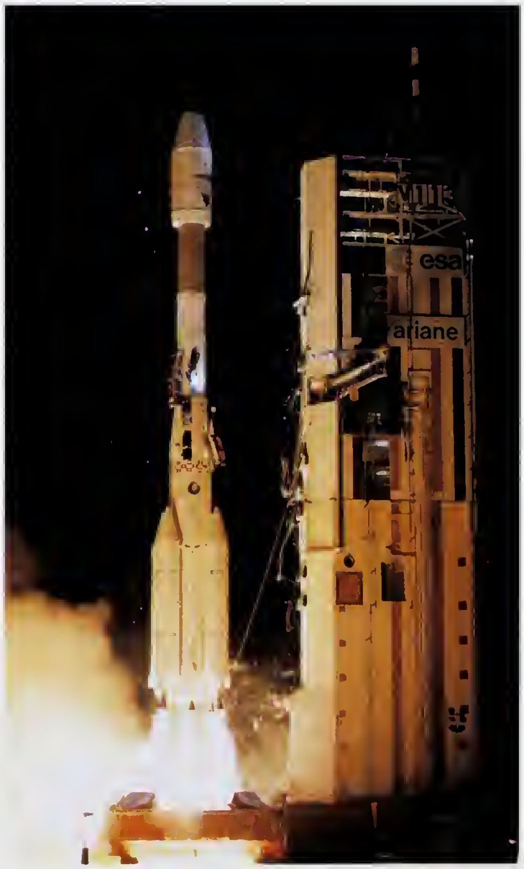
Plate 8 (Top) By courtesy of Lockheed Martin Corporation, (middle left) Sovfoto, (bottom left) courtesy of European Space Agency, (bottom right) AFP/Corbis



Soviet Molniya 1 communications satellite, one of a series deployed beginning in the mid-1960s.



First-generation Meteosat satellite, developed by the European Space Agency. Placed in geostationary orbits near the prime meridian, Meteosats have provided daily weather data for Europe, Africa, and the eastern Atlantic Ocean. The first Meteosat was launched on November 23, 1977.



Ariane-44L launch vehicle with two commercial communications satellites aboard, lifting off on October 28, 1998, from the European Space Agency's Kourou space centre in French Guiana.

slowly; the larger can be used while the ship is travelling at normal speeds. Plankton nets can be used to sample at one or more depths. Qualitative samplers sieve organisms from the water without measuring the volume of water passing through, whereas quantitative samplers measure the volume and hence the concentration of organisms in a unit volume of seawater.

The Clark-Bumpus sampler is a quantitative type designed to take an uncontaminated sample from any desired depth while simultaneously estimating the filtered volume of seawater. It is equipped with a flow meter that monitors the volume of seawater that passes through the net. A shutter opens and closes on demand from the surface, admitting water and spinning the impeller of the meter while catching the plankton. When the impeller is stopped by closing the shutter, the sampler can be raised without contamination from plankton in the waters above.

The midwater trawl is specially designed for rapid collection at depths well below the surface and at such a speed that active, fast-swimming fish are unable to escape from the net once caught. Trawls can be towed at speeds up to nine kilometres per hour. To counteract the tendency of an ordinary net to surface behind the towing vessel, a midwater trawl of the Isaacs-Kidd variety uses an inclined-plane surface rigged in front of the net entrance to act as a depressor. The trawl is shaped like an asymmetrical cone with a pentagonal mouth opening and a round closed end. Within the net, additional netting is attached as lining. A steel ring is fastened at the end of the net to maintain shape. A large perforated can is fastened by drawstrings on the end of the net to retain the sample in relatively undamaged condition.

The use of acoustics to record and measure the distribution of biological organisms is becoming a widely adopted practice. Some organisms can be tracked directly by their distinctive sounds. By recording and analyzing these sounds, biologists are able to chart the behaviour and distribution of such life-forms.

Organisms that passively affect various electronic systems are large mammals, schools of fish, and plankton that either scatter sound and so appear as false targets or background reverberation, or that attenuate the acoustic signal. Some fishes and invertebrates make up layers of acoustic-scattering material, which may exhibit daily vertical movement related to daily changes in light.

Light in the upper layers of the ocean is crucial to maintaining marine life. The penetration and absorption of light and the colour and transparency of the ocean water are indicative of biological activity and of suspended material. In situ measurements of water transparency and absorption include the submarine photometer, the hydrophotometer, and the Secchi disk. The submarine photometer records directly to depths of about 150 metres the infrared, visible, and ultraviolet portions of the spectrum. The hydrophotometer has a self-contained light source that allows greater latitude in observation because it can be used at any time of night or day and measures finer gradations of transparency. The Secchi disk, designed to measure water transparency, is a circular white disk that is lowered on a cable into the sea. In practice, the depth at which it is barely visible is noted. The greater the depth reading, the more transparent is the water.

The primary productivity of the ocean, which occurs in the upper layers, can be monitored by continuous measurement of absorption by chlorophyll molecules. This occurs in the red and blue portions of the spectrum, leaving the green to represent the characteristic colour of biological activity. Satellite measurements of ocean colour that span a number of wavelengths in the visible and infrared portions of the spectrum are used to give a large-scale view of the biological activity and suspended material in the ocean.

Exploration of the seafloor and the Earth's crust. The ocean floor has the same general character as the land areas of the world: mountains, plains, channels, canyons, exposed rocks, and sediment-covered areas. The lack of weathering and erosion in most areas, however, allows geological processes to be seen more clearly on the seafloor than on land. Undisturbed sediments, for example, contain a historical record of past climates and the state of the ocean, which has enabled geologists to find a close relation between past climates and the variation of the distance of the Earth from the Sun (the Milankovich effect).

Because electromagnetic radiation cannot penetrate any significant distance into the sea, the oceanographer uses acoustic signals, explosives, and earthquakes, as well as gravity and magnetic fields, to probe the seafloor and the structure beneath. Such techniques—which now include the capability to produce a swath, or two-dimensional, description of the seafloor beneath a ship—are providing increasingly accurate data on the shape of the ocean, its roughness, and the structure beneath. Satellite techniques are a more recent development. Because the shape of the sea surface is closely related to that of the seafloor due to gravity, satellite measurements of surface topography have been used to provide a global view of the ocean bottom. They also have provided data for an accurate mapping of such features as seamounts.

Marine sedimentation

Research on marine sedimentation involves the study of deposition, composition, and classification of organic and inorganic materials found on the seafloor. Samples of such materials are thoroughly examined aboard research vessels or in shore-based laboratories, where investigators analyze the size and shape of constituent particles, determine chemical properties such as pH, and identify and categorize the minerals and organisms present. From thousands of reported classifications and collected samples, bottom-sediment charts are prepared.

Various kinds of equipment are used to obtain samples from the seafloor. These include grabbing devices, dredges, and coring devices.

Grabbing devices, commonly known as snappers, vary widely in size and design. One general class of such devices is the clamshell snapper, which is used to obtain small samples of the superficial layers of bottom sediments. Clamshell snappers come in two basic varieties. One measures 76 centimetres in length, weighs roughly 27 kilograms (one kilogram = 2.2 pounds), and is constructed of stainless steel. The jaws of this device are closed by heavy arms, which are actuated by a strong spring and lead weight. It is capable of trapping about a pint of bottom material. The second type of clamshell snapper is appreciably smaller. Commonly called the mud snapper, this device is approximately 28 centimetres long and weighs 1.4 kilograms. Other grabbing devices include the orange peel bucket sampler, which is used for collecting bottom materials in shallow waters. A small hook attached to the end of the lowering wire supports the sampler as it is lowered and also holds the jaws open. When contact is made with the bottom, the sampler jaws sink into the sediment and the wire tension is released, allowing the hook to swing free of the sampler. Upon hoisting, the wire takes a strain on the closing line, which closes the jaws and traps a sample. The underway bottom sampler, or scoopfish, is designed to sample rapidly without stopping the ship. It is lowered to depths less than 200 metres from a ship moving at speeds no more than 28 kilometres per hour. The sampler weighs five kilograms and can capture samples ranging from mud to coral.

The second major category of bottom sampler is the dredge, which is dragged along the seafloor to collect materials. Bottom-dredging operations require very sturdy gear, particularly when dredging for rock samples. A typical dredge is constructed of steel plate and is 30 centimetres deep, 60 centimetres wide, and 90 centimetres long. The forward end is open, but the aft end has a heavy grill of round steel bars that is designed to retain large rock samples. When finer sized material is sought, a screen of heavy hardware cloth is placed over the grill.

Coring devices

Coring devices typically have three principal components: interchangeable core tubes, a main body of streamlined lead weights, and a tailfin assembly that directs the corer in a vertical line to the ocean bottom. The amount of sediment collected depends on the length of the corer, the size of the main weight, and the penetrability of the bottom. One type of coring device, the lightweight Phleger corer, takes samples only of the upper layer of the ocean bottom to a depth of about one metre. Deeper cores are taken by the piston corer. In this device, a closely fitted piston attached to the end of the lowering cable is installed inside the coring tube. When the coring tube is driven into the ocean floor, friction exerts a downward pull on the core sample. The hydrostatic pressure on the ocean bottom, however, exerts an upward pressure on the core that will work against a vacuum being created between the piston and the top of the core. The piston, in effect, provides a suction that overcomes the frictional forces acting between the sediment sample and the inside of the coring tube. The complete assembly of a typical piston core weighs about 180 kilograms and can be used to obtain samples as long as 20 metres. An improved version of this device, the hydraulic piston corer, is used by deep-sea drilling ships such as the "JOIDES Resolution." Essentially undisturbed cores of lengths up to 200 metres have been obtained with this type of corer.

Investigators may also make use of wire-line logging tools that are capable of measuring electrical resistance, acoustic properties, and magnetic and gravitational effects in the holes drilled. The "JOIDES Resolution" is equipped with tools of this sort, including a remote television camera, which are lowered into a drill hole after the core has been removed. Such wire-line logging apparatus make data immediately available for scientific analysis and decision making.

Use of acoustic techniques for studying the seafloor

Acoustic techniques have reached a high level of sophistication for geological and geophysical studies. Such multifrequency techniques as those that employ Seabeam and Gloria (Geological Long-Range Inclined Asdic) permit mapping two-dimensional swaths with great accuracy from a single ship. These methods are widely used to ascertain the major features of the seafloor. The Gloria system, for example, can produce a picture of the morphology of a region at a rate of up to 1,000 square kilometres per hour. Techniques of this kind are employed in conjunction with seismic reflection techniques, which involve the use of multichannel receiving arrays to detect sound waves triggered by explosive shots (*e.g.*, dynamite blasts) that are reflected off of interfaces separating rocks of different physical properties. Such techniques make it possible to measure the structure of the Earth's crust deep below the seafloor. Figure 8 shows the geometry for two ships engaged in acoustic profiling of the terrestrial crust. The depth to which the crust can be surveyed corresponds roughly to the spacing between the ships—in this particular case, nearly 10 kilometres. (A.B.R./D.J.Ba.)

By courtesy of the Institute for Geophysics, University of Texas at Austin

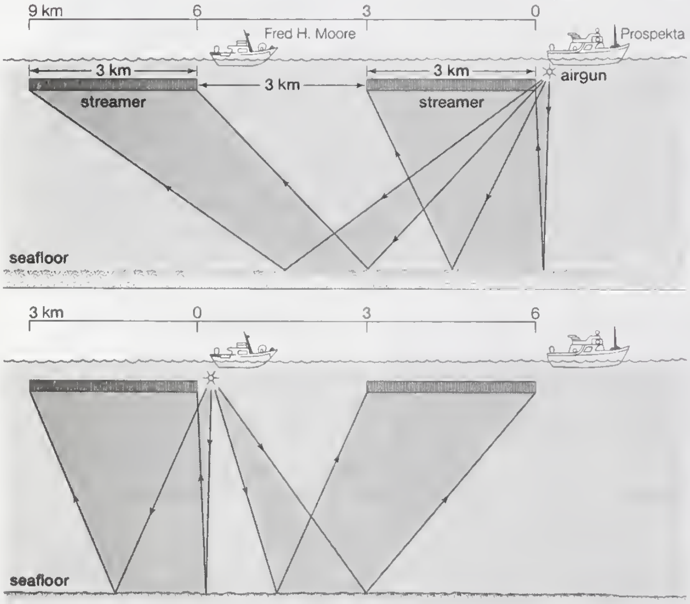


Figure 8: *Geometry for seismic profiling by two ships seen in a two-shot sequence.*
(Top) The lead ship shooting and shots recorded on its streamer and that of the trailing ship. (Bottom) The trailing ship shooting and shots recorded on its streamer and that of the lead ship.

Space exploration

Humans have always looked at the heavens and wondered about the nature of the objects that they could see in the night sky. With the development of rockets and the advances in electronics and other technologies in the 20th century, it became possible to send machines and then people above Earth's atmosphere into outer space. Achieving spaceflight enabled humans to begin to explore the solar system and the rest of the universe, to understand the many objects and phenomena that are better observed from a space perspective, and to use for human benefit the resources and attributes of the space environment. All of these activities—discovery, scientific understanding, and the application of that understanding to serve human purposes—are elements of space exploration.

ELEMENTS OF SPACEFLIGHT

Definition of space

Space, as considered here, is defined as all the reaches of the universe beyond Earth's atmosphere. There is no definitive boundary above Earth at which space begins, but in terms of the limiting altitude for vehicles designed for atmospheric flight, it may be considered to be as low as 45 kilometres (28 miles). The lowest practical orbit for an artificial satellite around Earth is about 160 kilometres (100 miles).

The space that separates cosmic objects is not entirely empty. Throughout this void, matter is scattered at extremely low densities. Nevertheless, space constitutes a much greater vacuum than has been achieved on Earth. Additionally, space is permeated by gravitational and magnetic fields, a wide spectrum of electromagnetic radiation, and high-energy cosmic-ray particles.

Kinds of spacecraft. *Spacecraft* is a general term that includes sounding rockets, unmanned artificial satellites and space probes, space stations, and vehicles for carrying humans to and from space. With the exceptions of the sounding rocket and the space shuttle, spacecraft are considered separately from the rocket-powered vehicle that launches the spacecraft into orbit or boosts it away from Earth's vicinity. A space probe is an unmanned spacecraft that is given a velocity great enough to allow it to escape Earth's gravitational attraction. A deep-space probe is a probe sent beyond the Earth-Moon system; if sent to explore other planets, it is also called a planetary probe. A space station is an artificial structure placed in orbit and equipped to support human habitation for extended periods.

Spacecraft differ greatly in size, shape, complexity, and purpose. Lightness of weight and functional reliability are primary features of spacecraft design. Depending on their mission, spacecraft may spend minutes, days, months, or years in the environment of space. Mission functions must be performed while exposed to high vacuum, extreme variations in temperature, and strong radiation.

A general differentiation of spacecraft is by function—scientific or applications. A scientific satellite or probe carries instruments to obtain data on magnetic fields, space radiation, the Sun or other stars, planets and their moons, and other astronomical objects and phenomena. Applications spacecraft have utilitarian tasks; examples are Earth observation, military reconnaissance, telecommunications, and navigation and global positioning satellites.

Launching into space. *Gravity.* Earth's gravitational attraction was one of the major obstacles to spaceflight. Owing to the observations and calculations of earlier scientists, rocket pioneers understood Isaac Newton's laws of motion and other principles of spaceflight, but the application of those principles had to await the development of rocket power to launch a spacecraft to the altitude and velocity required for its mission.

A spacecraft and its launch vehicle are projected upward as a reaction to the high-speed jet of combustion gases produced by the vehicle's rocket engines. If, for example, the total lifting force, or thrust, of the engines is twice the weight of the entire spacecraft-vehicle assembly at liftoff, then the assembly will rise at an initial acceleration equal to the standard gravitational acceleration (abbreviated *g*) of 9.8 metres (32 feet) per second per second. As propellant mass is consumed and ejected from the rocket engines, the vehicle continually lightens and acceleration increases.

Change in gravity with altitude

Earth's gravitational pull on the rising spacecraft subsides slowly. At an altitude of 160 kilometres (100 miles), it is still 95 percent of that at Earth's surface, and at 2,700 kilometres (1,680 miles) it is 50 percent (4.9 metres per second per second). For the purpose of spaceflight, the gravitational pull of Earth becomes negligible only at distances of several million kilometres, except when a spacecraft approaches the Moon and lunar gravity (one-sixth that of Earth) becomes predominant.

Staging. Staging describes a technique in which two or more rocket systems are mounted in vertical sequence, forming a multistage launch vehicle. Initially, the lowest, or first, stage is ignited, and it (sometimes assisted by attached booster rockets) lifts the vehicle at increasing velocity until its propellants are exhausted. At that point the stage drops off, lightening the vehicle, and the second stage is ignited. This stage, which is smaller and of lower thrust, then accelerates the launch vehicle further. The use of additional stages generally follows the same pattern.

For some missions the final stage is not employed during the initial climb into space but reserved for a later step of the flight. For example, a spacecraft carried on a three-stage vehicle may use the first two stages to achieve a low "parking orbit" around Earth. It is then boosted to a higher orbit or away from Earth by the third stage.

Acceleration rates. In general, the longer it takes a space vehicle to leave Earth's atmosphere and achieve required velocity, the less economical the procedure becomes. At low accelerations the launch vehicle wastes great quantities of propellant because, in effect, it is investing nearly 10 metres per second of velocity cach second of travel just to counter Earth's gravitational acceleration. An upper limit of acceleration is governed by the maximum accelerative stress permissible upon the vehiclc structure and the maximum spacecraft payload. In manned spaceflight an acceleration about six times that of gravity (6 *g*) is considered the maximum tolerable when the human body is positioned perpendicular to thc acceleration force—*i.e.,* with the head and heart at the same level.

Maximum acceleration for humans

Flight trajectories. There are four general types of trajectories: sounding rocket, Earth orbit, Earth escapc, and planetary.

Sounding rocket. Sounding rockets provide the only means of making scientific measurements at altitudes of 45–160 kilometres (28–100 miles), betwecn the maximum altitude of balloons and the minimum altitude of orbiting satellites. They can be single-stage or multistage vchicles and are launched nearly vertically. After all of the rockct stages have expended their fuel and dropped away, the payload section continues to coast upward, slowly losing speed because of gravity. Upward velocity drops to zero at peak altitude, and the payload then begins to fall. Typically, the payload is retrieved by parachute and flown again. Prior to parachute deployment, the flight path follows a parabolic trajectory, and flight time is less than 30 minutes.

Earth orbit. Flight into Earth orbit is achieved by launching a rocket vertically and then tilting its trajectory so that its flight is parallel to Earth's surface at the time that its final stage reaches orbital velocity at the desircd altitude. At this precise point, the rocket engine is shut down. A spacecraft carried aboard the rocket is then in free fall about Earth, the centrifugal pull on the spacecraft being equal to Earth's pull of gravity. At an altitude of 200 kilometres (125 miles), the velocity required to orbit Earth is about 29,000 kilometres (18,000 miles) per hour. Because this altitude is above most of the atmosphere, aerodynamic drag is not great, and the spacecraft will continue to orbit for an extended time.

The time required for a satellite to make one complete revolution is called the orbital period. At 200 kilometres, this is about 90 minutes. As the altitude increases, the orbital velocity of a satellite decreases and the orbital period increases. For example, at an altitude of 1,730 kilometres (1,075 miles), the orbital velocity is 25,400 kilometres (15,780 miles) per hour, and the period is two hours.

Geostationary orbit

At about 35,800 kilometres (22,250 miles), a satellite's velocity is 11,100 kilometres (6,900 miles) per hour, and its orbital period has a special value. It is equal to a sidereal day, the rotational period of Earth measured against the fixed stars (about four minutes shorter than the conventional 24-hour solar day). A satellite in this orbit has properties desirable for certain applications. For example, if the orbit lies in the plane of Earth's Equator, the satellite appears to be stationary in the sky. This particular orbit, called a geostationary orbit, is used for communications and meteorological satellites.

All of the above figures assume a circular orbit, which for a satellite is often ideal but difficult to achieve. Usually a satellite's orbit is an ellipse with a perigee altitude (nearest distance to Earth) and an apogee altitude (farthest distance from Earth). If thrust is available, a satellite's orbit may be made more nearly circular by reducing the velocity at perigee (which lowers the apogee) or by increasing the velocity at apogee (which raises the perigee). Thrust in such instances is applied against or in the direction of flight, respectively.

In launching a satellite into Earth orbit, the launch vehicle most commonly is tilted after liftoff in an easterly direction. Launching to the east is done to take advantage of the velocity imparted to the vehicle by Earth's eastward rotation. This rotational surface velocity is about 1,670 kilometres (1,037 miles) per hour at the Equator and 1,470 kilometres (913 miles) per hour at the latitude of Cape Canaveral, Florida. At the higher latitude of Russia's Baikonur launch site in Kazakhstan, the surface velocity is 1,170 kilometres (727 miles) pcr hour. It would be possible to launch a satellitc into a westerly orbit, but additional velocity, and thus additional fuel expenditure, would be required to achieve an orbit of the same altitude compared with an easterly orbit.

Polar orbit

If the satellite is to be put into a polar orbit—an orbit that crosses over Earth's poles—it is launched in a northerly or southerly direction. Although the bcnefit of an eastcrly launch is lost, a satellite in an orbit pcrpcndicular to the Equator offers other advantages. As Earth turns on its axis, the satellite travels over all parts of the globc cvery fcw revolutions. Satcllitcs that monitor Earth's environmcnt, such as remote sensing satellites and some weather satcllites, use polar orbits.

Earth escape. In order to escapc completely from Earth's gravity, a spacecraft requires a launch vclocity of about 40,000 kilometres (25,000 miles) per hour. If it subsequently docs not come undcr thc gravitational influencc of anothcr cclestial body, it will go into an orbit about thc Sun like a tiny planetoid. With precisc timing, a spaccraft can be sent on a trajectory that will carry it ncar thc Moon. In thc casc of thc Apollo lunar landing flights, the spacecraft was placed on a trajectory calculated to pass ahead of the Moon and, under the influcnce of lunar gravity, to swing around thc far side. If no velocity-changing maneuver had been made, the spacccraft would have looped around the Moon and rcturned on a trajectory toward Earth. By reducing flight spccd on the Moon's far side, Apollo astronauts placed thcir craft in a lunar orbit held by lunar gravity.

The so-called three-body problem of celestial mechanics (in the case of the Apollo missions, the relativc motions of Earth, the spacecraft, and the Moon undcr their mutual gravitational influence) is extremely complcx and has no general solution. Although equations exprcssing the relative motions can be written for specific cascs, no expcdicnt approximate solutions were possible before the development of high-speed digital computcrs for calculating trajectories of long-range missiles. Computcrs intcgratc the complicated equations of motion numerically, show thc spacecraft's complete trajectory at succcssivc positions through space, and compare the actual flight path to the preplanned path at any point in time.

Planetary flights. Because of the elliptical nature of planetary orbits, distances vary between Earth and the other planets. In the case of Earth's nearest neighbours, Venus and Mars, a so-called favourable launch opportunity occurs about every two years. Flights can be made at other times, but the velocity required is greater and length of time is longer, or, for a given launch vehicle, the payload must be lighter in weight.

The trajectory from Earth to Venus or Mars can be planned to take advantage of the changing orbital rela-

Hohmann, or transfer, orbit

tionships of the planets for the most economical flight in terms of fuel and energy. Such advantageous paths, called Hohmann or transfer orbits, were described in the 1920s. Although these trajectories require the least velocity, they are of long duration—as long as 260 days to Mars, for example. Thus, a compromise trajectory is often used, as in the case of Mariners 6 and 7 in 1969. Launched on February 24, 1969, Mariner 6 passed within 3,430 kilometres (2,130 miles) of the planet 157 days later, when Mars was 92.8 million kilometres (57.7 million miles) from Earth.

Gravity-assist technique

Some trajectories use the fall into a planet's gravitational field to transfer momentum from the planet to the spacecraft, thereby increasing its velocity and altering its direction. This gravity-assist, or slingshot, technique has been used by numerous planetary probes. For example, the Galileo probe during its six-year voyage to Jupiter swung by Venus once and Earth twice in order to reach its ultimate target in 1995.

The same considerations for planetary trajectories apply to spacecraft destined for other objects in deep space, such as asteroids and comets. For instance, the flight path of the Near Earth Asteroid Rendezvous (NEAR) spacecraft, which reached the asteroid Eros in 2000, incorporated a trajectory-reshaping flyby of Earth.

Trajectory correction

Navigation, docking, and recovery. *Navigation.* Traveling from point A to point B in space is almost never in a straight line or at constant velocity because of the many influences on the body in motion. The basis for space navigation is inertial guidance—*i.e.,* guidance based on the inertia of a spinning gyroscope, irrespective of external forces and without reference to the Sun or stars. By the use of three gyroscopes and accelerometers, a spacecraft's navigation system can make precise measurements of any change in velocity, either positive or negative, along any or all of the three principal axes. By changing attitude (conducting rotation about one or more axes) and firing one or more thrust motors, a spacecraft can make corrections to its trajectory. Preprogrammed computers, both on the ground and in larger spacecraft, continually monitor where the spacecraft is, where it was, and where it is supposed to be going.

During the launch phase, corrections to deviations in the planned flight path are usually made at once by small thrust motors on the launch vehicle, by deflection of the rocket exhaust jet, or by swinging one or more of the rocket engines in a gimbal mount. In the case of a rendezvous and docking between two spacecraft, radar data inform a crew—or, in the case of automated maneuvers, a computer—of the corrections required along each axis. With the implementation of the satellite-based Navstar Global Positioning System in the 1980s, it became possible for spacecraft to verify their locations within a few metres and their speeds within a few metres per second.

Rendezvous and docking. Rendezvous is the process of bringing two spacecraft together, whereas docking is their subsequent meeting and physical joining. The essential elements of a rendezvous are the matching of orbital trajectories and the movement of one spacecraft within close proximity of the other, typically within 100 metres (330 feet). Ideally, the two spacecraft also should lie in the same orbital plane.

Ordinarily for a rendezvous mission, one spacecraft is already in orbit, and the second spacecraft is launched to meet it. To achieve rendezvous, the launch of the second craft is timed within a fraction of a second. Because the orbiting spacecraft already has a high velocity relative to the second spacecraft on the ground, the second craft is launched well before the first passes overhead. The aim is to establish a coplanar orbit just below the first spacecraft. In this configuration the second craft, being at a lower orbit, is traveling at a faster speed and will overtake the first. When it is slightly ahead of the first spacecraft, it fires thrusters in a way that causes it to rise in orbit and thus to slow down until it matches the first craft's orbital altitude and velocity. Radar systems and on-board computers are necessary for such operations.

First rendezvous in space

Gemini 6 and 7 in 1965 were the first spacecraft to perform a rendezvous. In the Apollo lunar landing missions, the ascent stage of the Lunar Module rose from the Moon's surface to rendezvous and dock with the orbiting Command Module. Russian Soyuz spacecraft and U.S. space shuttle orbiters have rendezvoused and docked routinely with space stations. Whereas the United States has relied on human crews for close rendezvóus and docking, Russian spacecraft can perform these maneuvers automatically using technology developed and refined in the Soviet space program.

Because of payload limitations, spacecraft beyond a certain size and complexity cannot be launched into Earth orbit at one time. Building a large structure such as the International Space Station or, similarly, a future spacecraft for a human trip to Mars requires reliable rendezvous and docking techniques that can be used to assemble component parts taken to orbit on separate launches. Furthermore, rotation of space crews and emergency rescue missions require rendezvous and docking capability.

Reentry and recovery. Reentry refers to the return of a spacecraft into Earth's atmosphere. The blanket of relatively dense gas surrounding Earth is useful as a braking, or retarding, force resulting from aerodynamic drag. A concomitant effect, however, is severe frictional heating caused by rapid flow of atmospheric gas molecules along the spacecraft's blunt forward profile. Initially, heat shields were made of ablative materials that carried away the heat of reentry as they were shed, but the space shuttle introduced refractory materials—silica tiles and a reinforced carbon-carbon material—that withstood the heat directly.

Angle of reentry

Inherent in the safe reentry of a spacecraft is precise control of the angle of reentry. This angle with respect to Earth's horizon is −6.2° and is held within limits of 1° in either direction. If the reentry angle is too shallow, the spacecraft will skip or bounce off the atmosphere and back into space. If the angle is too great, the heat shield will not survive the extreme heating rates nor the spacecraft the high forces of deceleration. Returning Apollo Command Modules approached Earth at nearly 40,000 kilometres (25,000 miles) per hour. Even with a satisfactory reentry angle, the capsules' heat shields were subjected to temperatures approaching 3,000° C (5,400° F).

During the final phases of descent, a spacecraft typically deploys parachutes, which lower the vehicle to a soft landing. The Apollo Command Modules employed this technique to make ocean splashdowns. Russian Soyuz spacecraft traditionally soft-land on the ground. The reentry procedure of the winged space shuttle orbiter differs markedly: it descends by gliding and lands on a runway like an ordinary airplane. (F.C.D. III/Da.D./Ed.)

PRELUDE TO SPACEFLIGHT

From ancient times, people around the world have studied the heavens and used their observations and explanations of astronomical phenomena for both religious and practical purposes. Some even dreamed of leaving Earth to explore other worlds. In order to translate these visions of space travel into reality, it was necessary to devise some practical means of countering the influence of Earth's gravity. By the beginning of the 20th century, the centuries-old technology of rockets had advanced to the point at which it was reasonable to consider their use to accelerate objects to a velocity sufficient to enter orbit around Earth or even to escape Earth's gravity and travel away from the planet.

Pioneers in rocketry. *Tsiolkovsky.* The first person to study in detail the use of rockets for spaceflight was the Russian schoolteacher and mathematician Konstantin Tsiolkovsky. In 1903 his article "Exploration of Cosmic Space by Means of Reaction Devices" laid out many of the principles of spaceflight. Up to his death in 1935, Tsiolkovsky continued to publish sophisticated studies on the theoretical aspects of spaceflight. He never complemented his writings with practical experiments in rocketry, but his work greatly influenced later space and rocket research in the Soviet Union and Europe.

Goddard. In the United States, Robert Goddard became interested in space exploration after reading works such as H.G. Wells's *The War of the Worlds* (1898). He received his first two patents for rocket technology in 1914, and, with funding from the Smithsonian Institution, he published a theoretical treatise, *A Method of Reaching Ex-*

treme Altitudes, in 1919. Goddard's claim that rockets could be used to send objects as far as the Moon was widely ridiculed in the public press. Thereafter, the already shy Goddard conducted much of his work in secret, preferring to patent rather than publish his results. This approach limited his influence on the development of American rocketry, although early rocket developers in Germany took notice of his work.

Goddard's first rocket launch

In the 1920s, as a professor of physics at Clark University in Worcester, Massachusetts, Goddard began to experiment with liquid-fueled rockets. His first rocket, launched in Auburn, Massachusetts, on March 16, 1926, rose 12.5 metres (41 feet) and traveled 56 metres (184 feet) from its launching place. The noisy character of his experiments made it difficult for Goddard to continue work in Massachusetts. With support from aviator Charles Lindbergh and financial assistance from the philanthropic Daniel Guggenheim Fund for the Promotion of Aeronautics, he moved to Roswell, New Mexico, where from 1930 to 1941 he built engines and launched rockets of increasing complexity.

Oberth. The third widely recognized pioneer of rocketry, Hermann Oberth, was by birth a Romanian but by nationality a German. Reading Jules Verne's *From the Earth to the Moon* (1865) as a youth inspired him to study the requirements for interplanetary travel. Oberth's 1922 doctoral dissertation on rocket-powered flight was rejected by the University of Heidelberg for being too speculative, but it became the basis for his classic 1923 book, *Die Rakete zu den Planetenräumen* ("The Rocket into Interplanetary Space"). The work explained the mathematical theory of rocketry, applied the theory to rocket design, and discussed the possibility of constructing space stations and of traveling to other planets.

In 1929 Oberth published a second influential book, *Wege Zur Raumschiffahrt* (*Ways to Spaceflight*). His works led to the creation of a number of rocket clubs in Germany, as enthusiasts tried to turn Oberth's ideas into practical devices. The most important of these groups historically was the Verein für Raumschiffahrt (VfR; "Society for Spaceship Travel"), which had as a member the young Wernher von Braun. Although his work was crucial in stimulating the development of rocketry in Germany, Oberth himself had only a limited role in that development. Alone among the rocket pioneers, Oberth lived to see his ideas become reality: he was Braun's guest at the July 16, 1969, launch of Apollo 11.

Early rocket development. *Germany.* It was space exploration that motivated the members of the German VfR to build their rockets, but in the early 1930s their work came to the attention of the German military. In 1932, Braun, at age 20, became chief engineer of a rocket development team for the German army. After Adolf Hitler came to power in 1933, Braun was named the civilian head of that team, under the military command of Walter Dornberger. To give Braun's engineers the needed space and secrecy for their work, the German government erected a development and test centre at Peenemünde on the coast of the Baltic Sea. There they developed, among other devices, the V-2 ballistic missile. Although built as a weapon of war, the V-2 later served as the predecessor of many of the rockets used in the early space programs of the United States and the Soviet Union.

Role of V-2 in early space programs

As World War II neared its end in early 1945, Braun and many of his associates chose to surrender to the United States, where they believed they would likely receive support for their rocket research and space exploration plans. Later in the year they were brought to the United States, as were their engineering plans and the parts needed to construct a number of V-2s. The German rocket team played a central role in the early development of space launchers for the United States.

United States. In 1936, as Braun was developing rockets for the German military, several young American engineers led by graduate student Frank Malina began working on rocketry at the Guggenheim Aeronautical Laboratory of the California Institute of Technology (GALCIT). In 1943 Malina and his associates began calling their group the Jet Propulsion Laboratory (JPL), a name that was formally adopted the following year. JPL soon became a centre for missile research and development for the U.S. Army. Following World War II those weapons were adapted for use in early U.S. space experiments. After 1958, when it became part of the newly established National Aeronautics and Space Administration (NASA), JPL adapted itself to be the leading U.S. centre for solar system exploration.

Soviet Union. In the U.S.S.R., the government took an interest in rockets as early as 1921 with the founding of a military facility devoted to rocket research. Over the next decade that centre was expanded and renamed the Gas Dynamics Laboratory. There in the early 1930s Valentin Glushko carried out pioneering work on rocket engines. Meanwhile, other rocket enthusiasts in the Soviet Union organized into societies that by 1931 had consolidated into an organization known as GIRD (the abbreviation in Russian for "Group for the Study of Reactive Motion"), with branches in Moscow and Leningrad. Emerging as a leader of the Moscow branch was the aeronautical engineer Sergey Korolyov, who had become interested in spaceflight at a young age. Korolyov and a colleague, Mikhail Tikhonravov, on August 17, 1933, launched the first Soviet liquid-fueled rocket. Later that year, the Moscow and Leningrad branches of GIRD were combined with the Gas Dynamics Laboratory to form the military-controlled Rocket Propulsion Research Institute (RNII), which five years later became Scientific-Research Institute 3 (NII-3). In its early years the organization did not work directly on space technology, but ultimately it played a central role in Soviet rocket development.

Launch of first Soviet liquid-fueled rocket

Korolyov was arrested in 1937 as part of the Soviet leader Joseph Stalin's great purges of intellectuals and sent to a Siberian prison. After Stalin recognized the imprudence of removing the best technical people from the Soviet war effort, Korolyov was transferred to a prison-based design bureau, where he spent most of World War II working on weapons, although not on large rockets. By the end of the war Stalin had become interested in ballistic missiles, and he sent a team, which included Korolyov, on visits to Germany to investigate the V-2 program. A number of German engineers were relocated to the Soviet Union in the aftermath of the war, but they did not play a central role in postwar Soviet rocket development.

ON THE BRINK OF SPACE

Between 1946 and 1951 the U.S. Army conducted test firings of captured German V-2 rockets at White Sands, New Mexico. These sounding-rocket flights reached high altitudes (120–200 kilometres [75–125 miles]) before falling back to Earth. Although the primary purpose of the tests was to advance rocket technology, the Army invited American scientists interested in high-altitude research to put experiments aboard the V-2s. An Upper Atmosphere Research Panel, chaired by the physicist James Van Allen, was formed to coordinate the scientific use of these rocket launchings. The panel had a central role in the early years of U.S. space science, which focused on experiments on solar and stellar ultraviolet radiation, the aurora, and the nature of the upper atmosphere. As the supply of V-2s dwindled, other U.S.-built sounding rockets such as the WAC Corporal, Aerobee, and Viking were put into use. In other countries, particularly the Soviet Union, rocket-based upper-atmosphere research also took place after World War II.

In the early 1950s, scientists began planning a coordinated international investigation of Earth, to be called the International Geophysical Year (IGY), that would be held in 1957–58 under the auspices of the International Council of Scientific Unions. By this time progress in rocket development had advanced such that orbiting of an artificial satellite around Earth by 1957 seemed feasible. At the urging of American scientists, IGY planners in 1954 called for scientifically instrumented satellites to be launched as part of IGY activities. Soon thereafter, the governments of the Soviet Union and the United States governments each announced plans to do so.

In the years following World War II, the United States and the U.S.S.R. became political and military competitors in what soon was being called the Cold War. Because the Soviet Union was a closed society, U.S. leaders gave high priority to developing technology that could help gather intelligence on military preparations within the Soviet borders. As orbiting satellites neared realization, the idea of equipping such satellites with cameras and flying them over Soviet territory became more attractive to U.S. planners, and the U.S. Air Force began work on a reconnaissance satellite project.

FROM SPUTNIK TO APOLLO

Sputnik 1

The first satellites. Although Soviet plans to orbit a satellite during the IGY had been discussed extensively in technical circles, the October 4, 1957, launch of Sputnik 1 came as a surprise, and even a shock, to most people. Prior to the launch, skepticism had been widespread that the U.S.S.R. possessed the technical capabilities to develop both a sophisticated scientific satellite and a rocket powerful enough to put it into orbit. Under Korolyov's direction, however, the Soviet Union had been building an intercontinental ballistic missile (ICBM), with engines designed by Glushko, that was capable of delivering a heavy nuclear warhead to American targets. That ICBM, called the R-7 or Semyorka ("Number 7"), was first successfully tested on August 21, 1957, which cleared the way for its use to launch a satellite. Fearing that development of the elaborate scientific satellite intended as the Soviet IGY contribution would keep the U.S.S.R. from being the first into space, Korolyov and his associates, particularly Tikhonravov, designed a much simpler 83.6-kilogram (184-pound) sphere carrying only two radio transmitters and four antennas. After the success of the R-7 in August, that satellite was rushed into production and became Sputnik 1. A second, larger satellite carrying scientific instruments and the dog Laika, the first living creature in orbit, was launched November 3. The even larger, instrumented spacecraft originally intended to be the first Soviet satellite went into orbit in May 1958 as Sputnik 3.

After President Dwight Eisenhower in May 1955 had committed the United States to an IGY satellite, the mission was assigned to the Naval Research Laboratory, rather than to the Army's Redstone Arsenal, where Braun worked, so that the work would not interfere with Redstone's higher-priority development of ballistic missiles. The Navy project, called Vanguard, would use a new launch vehicle based on modified sounding rockets to orbit a small scientific satellite. Vanguard made slow progress over the subsequent two years, but after Sputnik's success, the White House pressed to have the satellite launched as quickly as possible.

Explorer 1

Braun and his Army superiors had not agreed with the decision to assign the satellite mission to the Navy. After the launches of the first two Sputniks, they secured permission to attempt their own satellite launch. In anticipation of such a situation, they had kept in touch with JPL and Van Allen and so were able to prepare a satellite quickly. On January 31, 1958, Braun's Jupiter-C launch vehicle, a modified Redstone ballistic missile, carried into orbit Explorer 1, the first U.S. satellite. Designed at JPL, Explorer 1 carried Van Allen's experiment to measure cosmic rays. The results from this experiment and similar ones aboard other U.S. and Soviet satellites launched that same year revealed that Earth was surrounded by two zones of trapped particles, now known as the Van Allen radiation belts.

Corona reconnaissance satellites

Initial satellite launches were scientific in character, but U.S. government interest in reconnaissance satellites persisted. In February 1958, President Eisenhower authorized the development, under conditions of great secrecy, of such a spacecraft. The project, which came to be called Corona, would take pictures over the Soviet Union and return them to Earth by dropping the exposed film in a capsule that would be snatched out of the air as it parachuted back from space. After 12 failures, the first successful Corona mission took place on August 18, 1960; the returned film contained images of many previously unknown Soviet airfields and missile sites.

NASA

Development of space organizations. *United States.* As part of its response to the first Sputnik launches, the United States government debated how best to organize itself for its space activities. In February 1958 President Eisenhower created the Defense Advanced Research Projects Agency (DARPA) and assigned it responsibility for all U.S. space projects. Soon afterward, he decided to separate civilian from military space efforts and proposed the creation of a National Aeronautics and Space Administration to manage the civilian segment. After approval by Congress, NASA began operation on October 1, 1958. DARPA was not successful in establishing itself as a military space agency. By 1960, after the Army had been obliged to relinquish control of JPL and Braun's rocket team to NASA management, the U.S. Air Force had emerged as the leading military service for space.

Eisenhower also decided to create a separate organization to manage the secret reconnaissance satellite program. This effort resulted in the National Reconnaissance Office (NRO), jointly directed by the Department of Defense and the Central Intelligence Agency. The very existence of this organization was kept secret until 1992. The NRO operated the initial Corona program until 1972. It continued to manage the development of successor photointelligence satellite systems of increasing technological sophistication and also developed radar-surveillance and electronic-signals-collection satellites. All were operated under conditions of the highest secrecy.

After it received its mandate to send Americans to the Moon, NASA grew into a large organization. At the peak of the Apollo program, the agency had 34,000 employees; by the end of the 20th century, this labour force had shrunk to 19,000, but NASA remained by far the largest space agency in the world.

The U.S. Air Force had no separate organization for space until 1982, when the Air Force Space Command was created to manage its military space operations, which involved the use of satellites for meteorology, communication, navigation, and early warning of missile attack. The other U.S. military services soon created similar organizations to administer their smaller space activities. In 1985 these organizations were brought under a unified U.S. Space Command, dominated by the Air Force, which was responsible for 85 percent of military space activities.

Soviet Union. In contrast to the United States, the Soviet Union had no separate, publicly acknowledged space agency. For 35 years after Sputnik, various design bureaus—state-controlled organizations that actually conceived and developed aircraft and space systems—had great influence within the Soviet system. Rivalry among those bureaus and their heads, who were known as chief designers, was a constant reality and posed an obstacle to a coherent Soviet space program. Space policy decisions were made by the Politburo of the Central Committee of the Communist Party as well as the Soviet government's Council of Ministers. After 1965, the government's Ministry of General Machine Building was assigned responsibility for managing all Soviet space and missile programs; the Ministry of Defense was also quite influential in shaping space efforts. A separate military branch, the Strategic Missile Forces, was in charge of space launchers and strategic missiles. Various institutes of the Soviet Academy of Sciences, particularly the Institute for Space Research (IKI), proposed and managed scientific missions.

Russian Space Agency

Only after the dissolution of the U.S.S.R did Russia create a civilian organization for space activities. Formed in February 1992, the Russian Space Agency acted as a central focus for the country's space policy and programs. Although it began as a small organization, it quickly took on increasing responsibility for the management of nonmilitary space activities and, as an added charge, aviation efforts. It later was renamed the Russian Aviation and Space Agency.

Europe. In 1961, within four years of the launch of the first U.S. and Soviet satellites, the government of France created the French Space Agency (CNES), which grew to become the largest national organization of its kind in Europe. Gradually other European countries formed government or government-sponsored organizations for space, among them the German Aerospace Center (DLR), the British National Space Centre (BNSC), and the Italian

Space Agency (ASI). Still others included space as part of their science or technology ministries.

European Space Agency

In 1964, a European Space Research Organisation (ESRO), created at the initiative of European scientists to pool government resources in support of space science, began operations. Ten Western European countries and Australia joined the organization. In the same year, a parallel European Launcher Development Organisation (ELDO), which had seven European member states, was established to develop a space launch vehicle for Europe. In 1975, a new European Space Agency (ESA) was formed from ESRO and ELDO to carry out both of their tasks. At the beginning of the 21st century, ESA had 15 member states—Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Norway, The Netherlands, Portugal, Spain, Sweden, Switzerland, and the United Kingdom. Canada also participated in some ESA projects. With a budget that made it the world's second largest civilian space agency, ESA carried out a comprehensive program in space science, applications, and infrastructure development.

Japan. In Japan, the University of Tokyo created an Institute of Space and Astronautical Science (ISAS) in 1964. This small group undertook the development of scientific spacecraft and the vehicles needed to launch them, and it launched Japan's first satellite, Ōsumi, in 1970. In 1981, oversight of ISAS was transferred to the Japanese Ministry of Education. In 1969, the Japanese government founded a National Space Development Agency (NASDA), which subsequently undertook a comprehensive program of space technology and satellite development and built a large launch vehicle, called the H-II, for those satellites. In 2001, both ISAS and NASDA came under the control of the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Chinese rocket development

China. China's space program evolved largely in secret, under the joint control of the Chinese military and the Commission on Science, Technology, and Industry for the National Defense. The rocket engineer Qian Xuesen (Tsien Hsue-shen), who had worked at GALCIT in the 1940s and helped found JPL, returned to China after the Communist takeover of 1949, where he became the guiding figure in the development of Chinese missiles and launch vehicles, both originally derived from a Soviet ICBM. China developed a family of Long March boosters, which are used domestically and serve as competitors in the international commercial space launch market. Its space development has concentrated on applications such as communications satellites and Earth observation satellites for civilian and military use, and it has announced plans to initiate its own manned spaceflight program.

The first human spaceflights. During the 1950s, space planners in both the Soviet Union and the United States anticipated the launching of a human being into orbit as soon as the required launch vehicle and spacecraft could be developed and tested. Much of the initial thinking focused on some form of piloted space plane, which, after being launched atop a rocket, could maneuver in orbit and then return to Earth, gliding to a horizontal landing.

Ballistic capsule approach

Rather than base their human spaceflight programs on space planes, however, the Soviet Union and the United States, in their desire to put people into space as quickly as possible, opted for a less technically demanding ballistic approach. A person would ride in a capsule-like spacecraft atop a rocket to achieve orbit. At the end of the flight, another rocket (called a retro-rocket) would slow down the spacecraft enough for it to fall back to Earth on a ballistic trajectory. To accomplish this feat, the spacecraft would have to survive the intense heat caused by reentering the atmosphere at a high speed and then carry its passenger safely back to Earth's surface.

Vostok. Soon after the success of the first Sputniks, Korolyov and his associate Tikhonravov began work on the design of an orbital spacecraft that could be used for two purposes. One was to conduct photoreconnaissance missions and then return the exposed film to Earth. The other was to serve as a vehicle for the first human spaceflight missions. The spacecraft was known as Object K, but it was called Vostok when it was used to carry a human into space. Vostok had two sections—a spherical capsule in which the person would ride and a conical module that contained the instruments needed for its flight. The spacecraft was large for the time, weighing 4.73 metric tons. Vostok was designed so that the human aboard need not touch any control from launch to touchdown—he would be essentially just a passenger. Nor would he land with the spacecraft. Rather, he would be ejected from it at an altitude of seven kilometres (4.3 miles) and parachute to dry land, while the spacecraft landed nearby with its own parachutes.

Gagarin's historic spaceflight

After a series of five test flights carrying dogs and human dummies, the first human lifted into space in Vostok 1 atop a modified R-7 rocket on April 12, 1961, from the Soviet launch site at the Baikonur Cosmodrome in Kazakhstan. The passenger, Yury Gagarin, who was by that time being called a cosmonaut, was a 27-year-old Russian test pilot. After firing of the retro-rocket 78 minutes into the mission, the crew capsule separated from the instrument module—although not without problems—and Gagarin parachuted to a soft landing 108 minutes after his launch.

There were five additional one-person Vostok missions. In August 1961, Gherman Titov at age 25 (still as of 2001 the youngest person ever to fly in space) completed 17 orbits of Earth in Vostok 2. He became ill with space sickness (the equivalent of motion sickness on Earth) during the flight, an incident that caused a one-year delay in Vostok flights while Soviet physicians investigated the possibility that humans could not survive for extended times in the space environment. In August 1962, two Vostoks, 3 and 4, were orbited at the same time and came within 6.5 kilometres (four miles) of one another. This dual mission was repeatcd in June 1963; aboard the Vostok 6 spacecraft was Valentina Tereshkova, the first woman to fly in space.

Mercury. The initial U.S. effort to launch a human into space was known as Projcct Mercury. It was carried out by NASA, which had been given that responsibility ovcr Air Force objections. NASA engineers, led by Robert Gilruth and Maxime Faget, designed a small cone-shaped capsule for the mission. Compared with thc ncarly fivc-mctric-ton Vostok, it weighed 1.94 metric tons. Unlike the Soviet approach, in which a cosmonaut was orbited on the first human spaceflight, NASA planned several precautionary suborbital test flights in which an astronaut would bc in space only for a few minutes of his 15-minute up-and-down ride. The Mercury capsule would parachute with its passenger all the way back to Earth's surface, to land in the ocean and be recovered by Navy ships. Also in contrast to Vostok, the Mercury capsule was designed to allow the astronaut to control some aspccts of its flight whilc in spacc.

First U.S. orbital flights

The United States used chimpanzees, rather than dogs, as test subjects prior to human flights. On May 5, 1961, Alan Shepard made thc first manned Mercury flight atop a Redstone rocket. A second suborbital Mercury mission carrying Virgil Grissom followed in July. John Glenn became the first American astronaut to orbit Earth in his three-orbit mission on February 20, 1962. His Mercury spacecraft was launched by a modified Air Force Atlas ICBM. Three more one-man Mercury orbital flights, carrying astronauts Scott Carpenter, Walter Schirra, and L. Gordon Cooper, were conducted, the last being a 22-orbit mission in May 1963.

Gemini and Voskhod. In 1961, President John F. Kcnnedy announced that the United States would send people to the Moon "before this decade is out." In order to test many of the techniques that would be needed to carry out a lunar mission, particularly rendezvousing and docking two objects in space, the United States in late 1961 decided to develop a two-person spacecraft called Gemini. The Gemini spacecraft was much more complex than the rudimentary Mercury capsule and, at 3.81 metric tons, was twice as heavy. Another converted Air Force ICBM, a Titan II, was used to launch the Gemini spacecraft.

The first manned Gemini mission lifted into space in March 1965; nine more missions followed, the last in November 1966. On the second mission in June 1965, Edward White became the first American astronaut to operate outside a spacecraft. His 20-minute space walk—also known as extravehicular activity (EVA)—was without incident.

Although problems developed on many of the Gemini flights, the program demonstrated that people could live and work in space for as long as 14 days, more than the time needed for a round trip to the Moon. It also showed that astronauts could carry out rendezvous in space and could make useful observations of Earth.

As plans in the United States for multiple astronaut missions became known, the Soviet Union worked to maintain its lead in the space race by modifying the Vostok spacecraft so that it could carry as many as three persons. Korolyov could accomplish this only by having the crew fly without wearing spacesuits. The redesigned spacecraft was known as Voskhod. There were two Voskhod missions, one with three people aboard in October 1964 and another with a two-man crew in March 1965. On the second mission, cosmonaut Aleksey Leonov became the first human to leave an orbiting spacecraft, less than three months before White. His 12-minute EVA was full of problems, and his reentry of the Voskhod spacecraft was particularly difficult.

Soyuz. Korolyov and his associates began work in 1962 on a second-generation spacecraft, to be called Soyuz. It was to be a much more complex vehicle than Vostok, holding as many as three people in an orbital crew compartment, with a separate module for crew reentry and a third section containing spacecraft equipment and rocket engines for in-orbit and reentry maneuvers. Soyuz was to be capable not only of flights in Earth orbit but also, in modified versions, of flights around the Moon and even a lunar landing.

The first launch of Soyuz, with a single cosmonaut aboard, Vladimir Komarov, aboard, took place on April 23, 1967. Once the spacecraft reached orbit, it suffered a number of problems, which prompted ground controllers to bring Komarov back to Earth as soon as possible. After reentry, however, the spacecraft's main parachute did not fully deploy, and the Soyuz hit the ground at high speed. Komarov became the first person to perish during a spaceflight, and the accident dealt a major blow to Soviet hopes of orbiting or landing on the Moon before the United States.

First death during a spaceflight

After the problems with the Soyuz design were diagnosed and remedied, various models of the spacecraft served as the means of access to space for the Soviet, and then Russian, program of human spaceflight for more than 30 years. At the start of the 21st century, a version of Soyuz was used as the crew rescue vehicle—the lifeboat—for the early phase of construction and occupancy of the International Space Station.

The race to the Moon. *The American commitment.* In the immediate aftermath of Gagarin's orbital flight, President Kennedy was advised of Braun's belief that the Soviet Union, using Korolyov's existing R-7 launcher, could well succeed in sending a multiperson spacecraft into Earth orbit and perhaps even around the Moon before the United States. The first competition that the United States had a good chance of winning would be that of a manned lunar landing, because it would require each country to develop a new, more powerful rocket. On those technical grounds, and because a lunar landing would be a very visible demonstration of American strength, Kennedy announced on May 25, 1961, that the United States would commit itself to a lunar landing before 1970.

In response to Kennedy's decision, the United States carried out a warlike, but peaceful, mobilization of financial and human resources. NASA's budget was increased almost 500 percent in three years, and at its peak the lunar landing program involved more than 34,000 NASA employees and 375,000 employees of industrial and university contractors.

Elements of Apollo program

By the end of 1962 the basic elements of what was called Project Apollo were in place (see Figure 9). The launch vehicle would be a powerful Saturn V rocket, 110.6 metres (363 feet) tall and power driven by five huge engines generating a total of 33,000 kilonewtons (7.5 million pounds) of lifting power at takeoff—100 times the takeoff thrust of the Redstone rocket that had launched Shepard. After an intense debate, NASA chose a spacecraft configuration for Apollo that could be sent up in one launch, rather than a larger spacecraft that would need to be assembled in a series of rendezvous in Earth orbit. The Apollo spacecraft would have three sections. A Command Module would house the three-person crew on liftoff and landing and during the trip to and from the Moon. A Service Module would carry various equipment and the rocket engine needed to guide the spacecraft into lunar orbit and then send it back to Earth. A Lunar Module, comprising a descent stage and an ascent stage, would carry two people from lunar orbit to the Moon's surface and back to the Command Module. The ability of the Lunar Module's ascent stage to rendezvous and dock in lunar orbit with the Command Module after takeoff from the Moon was critical to the success of the mission.

The Soviet response. While committing the United States to winning the Moon race, President Kennedy also made several attempts in the early 1960s to convince the Soviet leadership that a cooperative lunar landing program between their two countries would be a better alternative. No positive reply from the Soviet Union was forthcoming, however. In fact, between 1961 and 1963 there was still vigorous debate within the Soviet Union over the wisdom of undertaking a lunar program, and no final decision had been made on the question.

Meanwhile, the separate Soviet design bureaus headed by Korolyov and Vladimir Chelomey competed fiercely for a lunar mission assignment, either a flight around the Moon or an actual landing. Finally, in August 1964, Korolyov received the lunar landing assignment, and, soon afterward, Chelomey was given responsibility for planning a circumlunar flight to be carried out before the 50th anniversary of the Bolshevik Revolution, which would take place in October 1967. In 1965, Soviet leaders decided to combine the efforts of the two rivals for the circumlunar mission, using a version of Korolyov's Soyuz spacecraft and a new rocket, the UR-500 (also called the Proton), designed by Chelomey.

Soviet manned lunar program

The rocket that Korolyov designed for the lunar landing effort was called the N1. Like the Saturn V, it was huge, standing 112.8 metres (370 feet) tall and having a planned takeoff thrust of 44,500 kilonewtons (10 million pounds). Instead of a few large rocket engines in its first stage, however, the N1 had 30 smaller engines. These were developed by Nikolay Kuznetsov, an aircraft-engine chief designer who had little experience with rocket engines, rather than the more capable Glushko. Korolyov and Glushko, already personal adversaries for many years, had disagreed on the proper fuel for the N1, and they finally decided that they could no longer work together.

Indecision, inefficiencies, inadequate budgets, and personal and organizational rivalries in the Soviet system thus posed major obstacles to success in the race to the Moon. To this was added the unexpected death on January 14, 1966, of Korolyov, at age 59, during surgery. His successor, Vasily Mishin, attempted to maintain the program's momentum, but he was not the effective manager or politically sophisticated operator that Korolyov had been.

Interim developments. In the United States, Apollo moved forward as a high-priority program. A major setback occurred on January 27, 1967, when astronauts Grissom, White, and Roger Chaffee were killed after their Apollo 1 Command Module caught fire during a ground test. The first manned Apollo mission, designated Apollo 7 and intended to test the redesigned Command Module, was launched into Earth orbit on October 11, 1968. The launcher used was a Saturn IB, a less-powerful rocket than the Saturn V needed to reach the Moon. The mission's success cleared the way for a bold step—the first launch of a crew atop a Saturn V to the lunar vicinity. On December 21, 1968, the Apollo 8 Command and Service modules, carrying astronauts Frank Borman, James Lovell, and William Anders, were put on a trajectory that sent them into orbit around the Moon on Christmas Eve, December 24.

First manned circumlunar voyage

One reason for conducting the Apollo 8 mission was to allow NASA to test most of the systems needed for a lunar landing attempt while waiting to carry out a manned trial in Earth orbit of the Lunar Module, whose development was behind schedule. Another was the concern that the Soviet Union would beat the United States in sending people

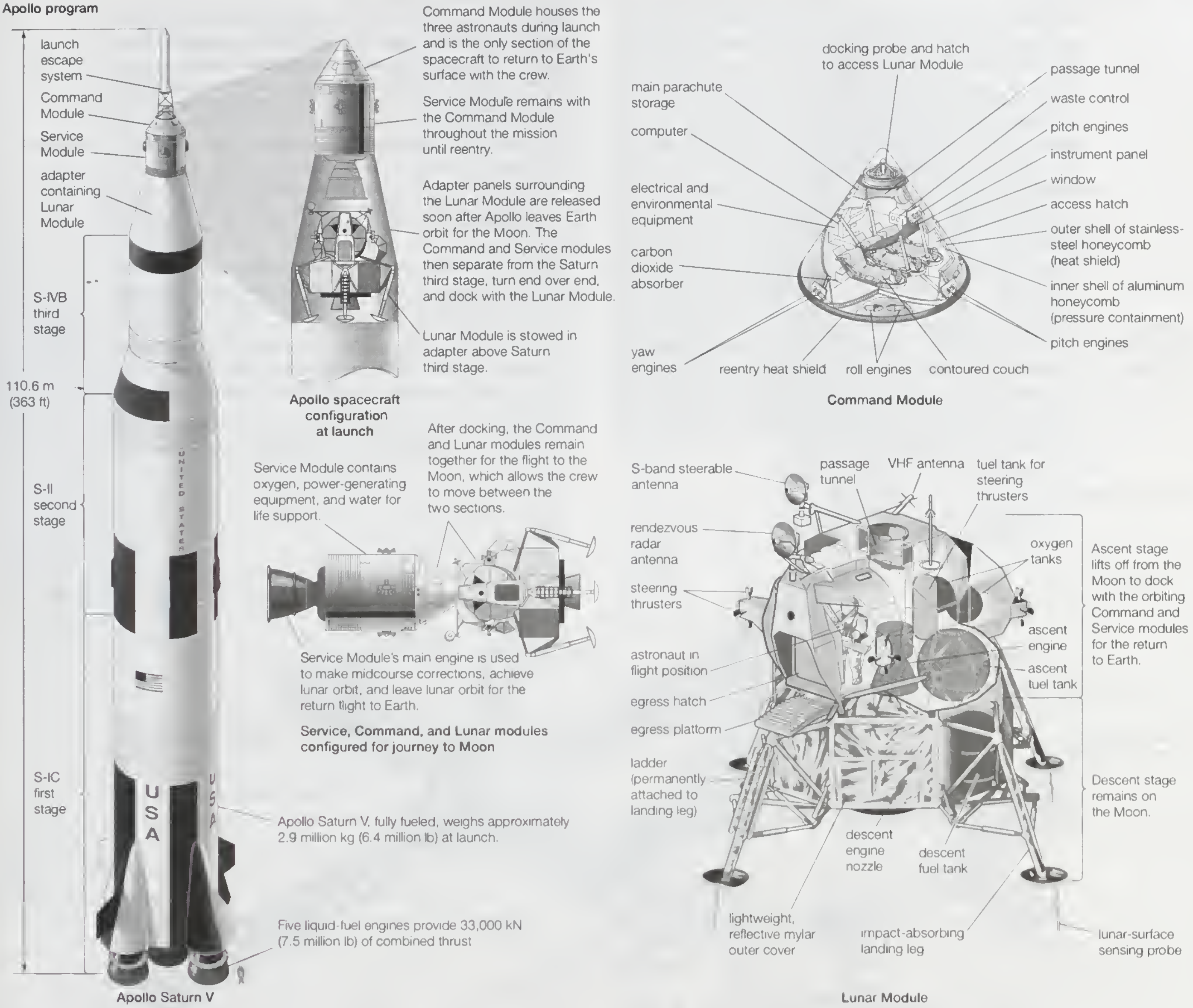

Apollo program
launch escape system
Command Module
Service Module
adapter containing Lunar Module
S-IVB third stage
110.6 m (363 ft)
S-II second stage
S-IC first stage
UNITED STATES
USA
USA
Apollo Saturn V
Command Module houses the three astronauts during launch and is the only section of the spacecraft to return to Earth's surface with the crew.
Service Modufe remains with the Command Module throughout the mission until reentry.
Adapter panels surrounding the Lunar Module are released soon after Apollo leaves Earth orbit for the Moon. The Command and Service modules then separate from the Saturn third stage, turn end over end, and dock with the Lunar Module.
Lunar Module is stowed in adapter above Saturn third stage.
Apollo spacecraft configuration at launch
Service Module contains oxygen, power-generating equipment, and water for life support.
After docking, the Command and Lunar modules remain together for the flight to the Moon, which allows the crew to move between the two sections.
Service Module's main engine is used to make midcourse corrections, achieve lunar orbit, and leave lunar orbit for the return flight to Earth.
Service, Command, and Lunar modules configured for journey to Moon
Apollo Saturn V, fully fueled, weighs approximately 2.9 million kg (6.4 million lb) at launch.
Five liquid-fuel engines provide 33,000 kN (7.5 million lb) of combined thrust
docking probe and hatch to access Lunar Module
main parachute storage
computer
electrical and environmental equipment
carbon dioxide absorber
yaw engines
reentry heat shield
roll engines
contoured couch
passage tunnel
waste control
pitch engines
instrument panel
window
access hatch
outer shell of stainless-steel honeycomb (heat shield)
inner shell of aluminum honeycomb (pressure containment)
pitch engines
Command Module
S-band steerable antenna
passage tunnel
VHF antenna
fuel tank for steering thrusters
rendezvous radar antenna
steering thrusters
oxygen tanks
ascent engine
ascent fuel tank
Ascent stage lifts off from the Moon to dock with the orbiting Command and Service modules for the return to Earth.
astronaut in flight position
egress hatch
egress platform
ladder (permanently attached to landing leg)
descent engine nozzle
descent fuel tank
Descent stage remains on the Moon.
lightweight, reflective mylar outer cover
impact-absorbing landing leg
lunar-surface sensing probe
Lunar Module

Figure 9: *The Apollo program.*
(Left) Saturn V launch vehicle, showing configuration of Command, Service, and Lunar modules at launch (top centre) and during the journey to the Moon (bottom centre).
(Right) Details of the Command Module (top) and Lunar Module (bottom).
Encyclopædia Britannica, Inc

to the lunar vicinity. A circumlunar mission indeed had been part of Soviet plans, but the Soyuz 1 accident had made the October 1967 deadline infeasible. During 1968, a number of test flights of a circumlunar mission were made using the Proton launcher and a version of the Soyuz spacecraft designated Zond. In September Zond 5 carried a biological payload, including two tortoises, around the Moon and safely back to Earth, but two months later the Zond 6 spacecraft depressurized and then crashed on landing, ending any hope for a quick follow-on launch with a human crew.

Failure of Soviet Moon rocket

The Soviet lunar landing program went forward rather fitfully after 1964. The missions were intended to employ the N1 launch vehicle and another variation of the Soyuz spacecraft, designated L3, that included a lunar landing module designed for one cosmonaut. Although an L3 spacecraft was constructed and three cosmonauts trained for its use, the N1 rocket was never successfully launched. After four failed attempts between 1969 and 1972, the N1 program was finally cancelled in May 1974, thus ending Soviet hopes for human missions to the Moon.

The Apollo lunar landings and Apollo-Soyuz. By contrast with the Soviet lunar landing efforts, during 1969 all went well for the Apollo program. In March the Apollo 9 crew successfully tested the Lunar Module in Earth orbit, and in May the Apollo 10 crew carried out a full dress rehearsal for the landing, coming within 15,200 metres (50,000 feet) of the lunar surface. On July 16, 1969, astronauts Neil Armstrong, Edwin ("Buzz") Aldrin, and Michael Collins set off on the Apollo 11 mission, the first lunar landing attempt. While Collins remained in lunar orbit in the Command Module, Armstrong piloted the Lunar Module, nicknamed Eagle, away from boulders on the lunar surface and to a successful landing on a flat lava plain called the Sea of Tranquillity at 4:18 PM U.S. Eastern Daylight Time on July 20. Six and a half hours later, Armstrong, soon followed by Aldrin, left the Lunar Module and took the first human step on the surface of another celestial body. As he did so, he noted, "That's one small step for [a] man, one giant leap for mankind." (In the excitement of the moment, Armstrong skipped the "a" in the statement he had prepared.) Concluding 2.5 hours of activity on the lunar surface, the two men returned to the Lunar Module with 21.7 kilograms (47.8 pounds) of lunar samples. Twelve hours later they blasted off the Moon in the Lunar Module's ascent stage and rejoined Collins in the Command Module. The crew returned to Earth on July 24, splashing down in the Pacific Ocean.

Humans on the Moon

The successful Apollo 12 mission followed in November 1969. The Apollo 13 mission, launched in April 1970, experienced an explosion of the oxygen tank in its Service Module on the outbound trip to the Moon. The crew survived this accident only through the improvised use of the Lunar Module as living quarters in order to preserve the remaining capabilities of the Command Module for reentering Earth's atmosphere after they had returned from their circumlunar journey. Four more Apollo missions followed. On the final three, the crew had a small cartlike rover that allowed them to travel several kilometres from their landing site. The final mission, Apollo 17, which was conducted in December 1972, included geologist Harrison Schmitt, the only trained scientist to set foot on the Moon.

Apollo-Soyuz Test Project

An Apollo spacecraft was used for the last time in 1975. Three years earlier, as a sign of improved U.S.-Soviet relations, the two countries had agreed to carry out a joint mission in which an Apollo spacecraft carrying three astronauts would dock in orbit with a Soyuz vehicle having two cosmonauts aboard. The Apollo-Soyuz Test Project, which took place in July 1975, featured a "handshake in space" between Apollo commander Thomas Stafford and Soyuz commander Aleksey Leonov.

ORBITING SPACE PLATFORMS

Salyut 1

Space stations. By 1969, even though the U.S.S.R. was still moving forward with its lunar landing program, it had begun to shift its emphasis in human spaceflight to the development of Earth-orbiting stations in which cosmonaut crews could carry out extended observations and experiments on missions that lasted weeks or months rather than a few days. The first Soviet space station, called Salyut 1, was launched April 19, 1971. Its initial crew, Georgy Dobrovolsky, Viktor Patsayev, and Vladislav Volkov, spent 23 days aboard the station carrying out scientific studies but perished when their Soyuz spacecraft depressurized during reentry.

Skylab

With similar objectives for a long-term manned platform in space, the United States converted the third stage of a Saturn V rocket into an orbital workshop for solar and biomedical studies. This first U.S. space station, called Skylab, was launched May 14, 1973. Over a period of eight and a half months, three three-person crews using Apollo spacecraft for transport spent time aboard Skylab, with the final crew staying for 84 days.

Whereas the United States did not launch a planned second Skylab due to budgetary cuts, the Soviet Union orbit-

Mir space station

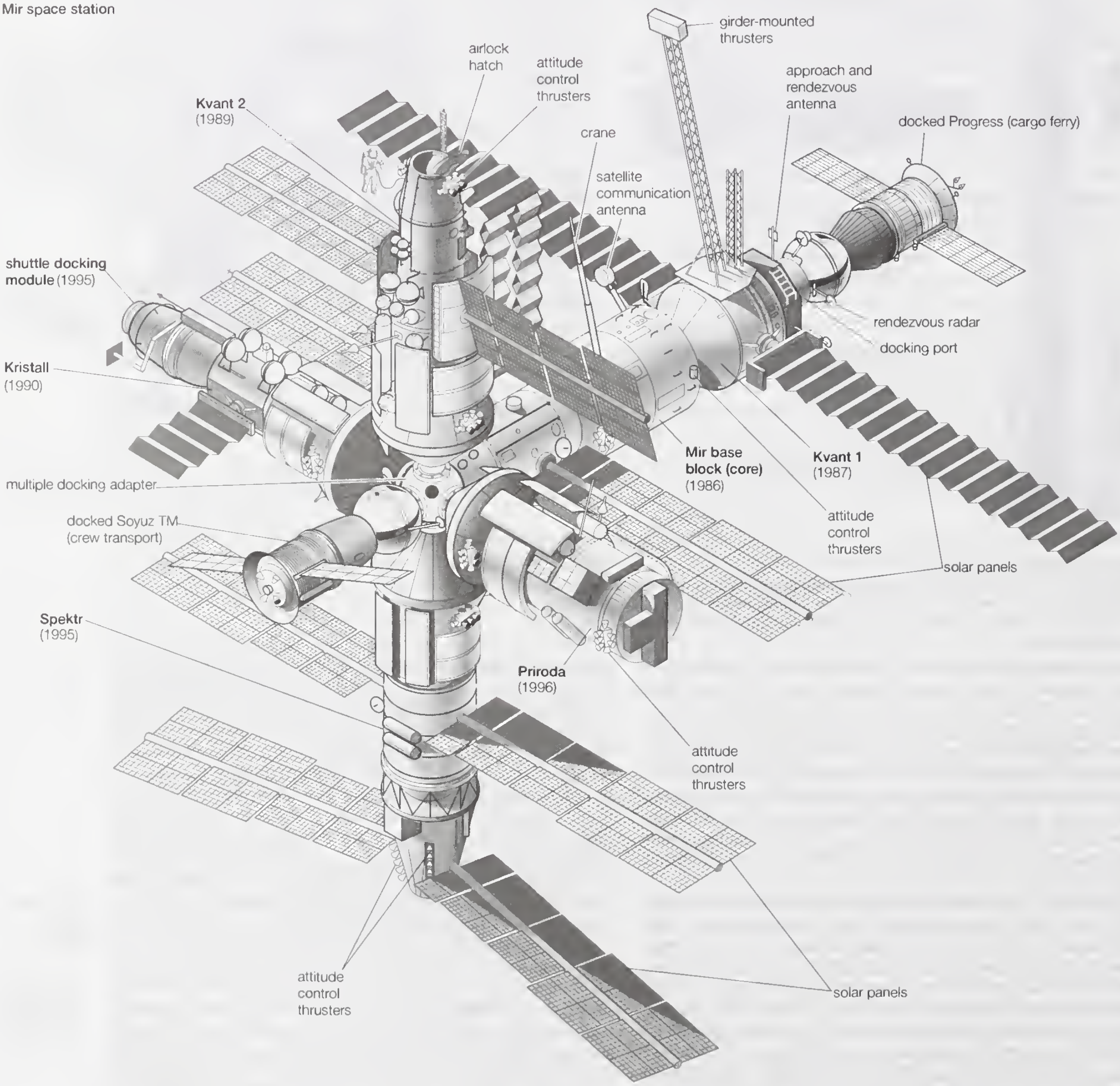

girder-mounted thrusters
airlock hatch
attitude control thrusters
approach and rendezvous antenna
Kvant 2 (1989)
docked Progress (cargo ferry)
crane
satellite communication antenna
shuttle docking module (1995)
rendezvous radar
docking port
Kristall (1990)
Mir base block (core) (1986)
Kvant 1 (1987)
multiple docking adapter
attitude control thrusters
docked Soyuz TM (crew transport)
solar panels
Spektr (1995)
Priroda (1996)
attitude control thrusters
attitude control thrusters
solar panels

Figure 10: *Soviet-Russian space station Mir, after completion in 1996.*
Date shown for each module is its year of launch. Docked to the station are a Soyuz TM manned spacecraft and an unmanned Progress resupply ferry.
Encyclopædia Britannica, Inc

ed and successfully occupied five more Salyut stations in a program that continued through the mid-1980s. Two of these stations had a military reconnaissance mission, but the others were devoted to scientific studies, particularly biomedical research. The Soviet Union also launched guest cosmonauts from allied countries for short stays aboard the Salyuts 6 and 7.

These early stations were a reflection of a long-held belief among space visionaries, dating back to Tsiolkovsky at the start of the 20th century, that living and working in space, first in Earth orbit and then on the Moon, Mars, and other locations, were an important part of the human future. It also was thought that increasingly complex orbital outposts would be the first steps in a long-term process of space development and colonization. The early focus of the United States and the U.S.S.R. on sending people to the Moon for political reasons deviated from this vision, which has since returned to dominate space thinking.

Mir

The Soviet Union followed its Salyut station series with the February 1986 launch of the core element of the modular Mir space station (Figure 10). Additional modules carrying scientific equipment and expanding the living space were attached to Mir in subsequent years. In 1994–95, Valery Polyakov, a medical doctor, spent 438 continuous days aboard the station. More than 100 different people from 12 countries visited Mir, including seven American astronauts in the 1995–98 period. The station, which was initially scheduled to operate for only five years, supported human habitation until mid-2000 (continuously between 1989 and 1999), although it experienced a number of accidents and other serious problems.

International Space Station

The United States did not follow up on Skylab until 1984, when President Ronald Reagan approved a space station program and invited U.S. allies to participate. By 1988, 11 countries—Canada, Japan, and nine countries from Europe—had decided to join what was known as Space Station Freedom. Progress in developing the station was slow, however, and in 1993, newly elected President Bill Clinton ordered a sweeping redesign of the program. The United States and its existing partners invited Russia, which had inherited most of the Soviet Union's space efforts after the U.S.S.R.'s collapse in 1991, to participate in the multinational program, renamed the International Space Station (ISS). Three additional countries joined during the 1990s, making the 16-country project the largest-ever cooperative technological undertaking. The first two elements of the ISS were launched and connected in space in late 1998, and several modules and other equipment were subsequently added. Assembly and initial operation of the facility were scheduled to take place over much of the first decade of the 21st century.

The space shuttle. After the success of the Apollo 11 mission, NASA proposed an ambitious plan that included human flights to Mars, a series of large space stations to be developed during the 1970s, and a new, reusable space transportation system to send people and supplies to those stations. This plan was quickly rejected, as there was no interest in major post-Apollo space programs among the political leadership or the general public. In 1972 NASA received presidential approval to develop a partially reusable transport vehicle called a space shuttle. This vehicle was intended to carry people and as much as 29,500 kilograms (65,000 pounds) of cargo into a low Earth orbit at low cost. Based on those expectations, the United States planned to use the shuttle as its sole launch vehicle once it entered operation and to operate a shuttle fleet with a launch rate as high as 60 per year. In the absence of a space station, plans also called for having the shuttle serve double duty as a space platform to conduct in-orbit research for periods as long as two weeks. To that end, Europe pledged to contribute a pressurized laboratory, known as Spacelab, that would be carried in the shuttle's payload bay.

The space shuttle design (Figure 11) had three major components. A reusable winged orbiter would carry crew and cargo and be able to glide to a landing on a conventional runway at the end of its mission. A large external tank would carry the liquid-oxygen and liquid-hydrogen propellants for the orbiter's three powerful engines. The tank would be used only during the first eight minutes of flight; once the fuel was exhausted, it would be discarded and burn up on reentry. Two solid-fuel rockets would assist in accelerating the vehicle during the first two minutes of flight; they would then be detached and parachute into the ocean, where they would be recovered for future use. A fleet of four operational orbiters, named *Columbia, Challenger, Atlantis,* and *Discovery,* was built.

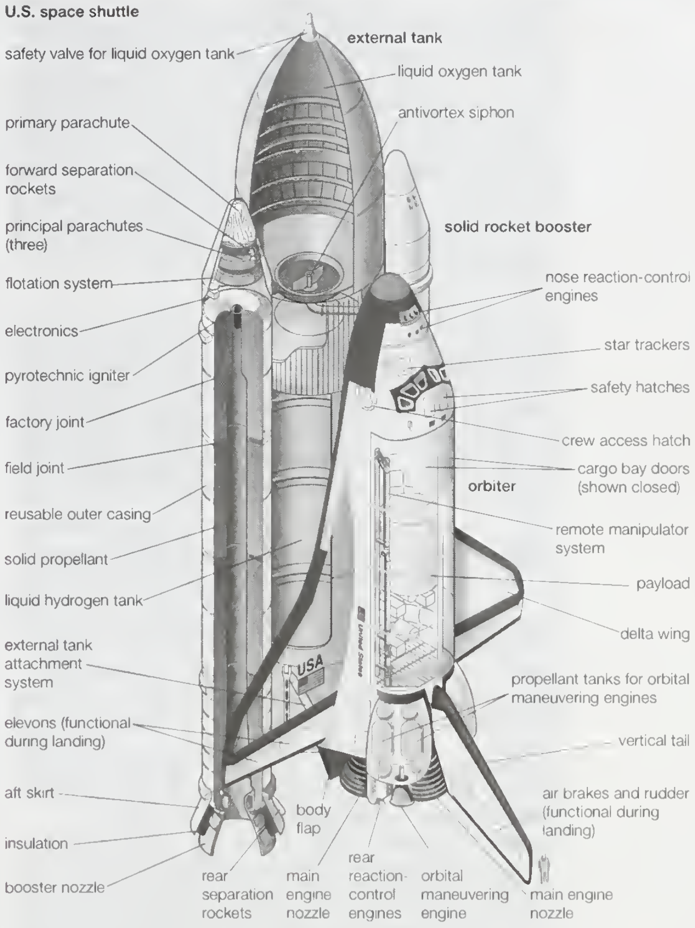


Figure 11: U.S. space shuttle, composed of a winged orbiter, an external liquid-propellant tank, and two solid-fuel rocket boosters.
Encyclopædia Britannica, Inc

After several years of technical and budgetary delays to the program, the first space shuttle flight took place on April 12, 1981; aboard were astronauts John Young, a veteran of the Gemini and Apollo programs, and Robert Crippen. With additional flights, it became evident that projections of the vehicle's operational costs and performance had been extremely optimistic. Major refurbishment was required between each launch; the highest flight rate achieved was in 1985, when the shuttle was launched nine times. Each launch cost hundreds of millions of dollars, rather than the tens of millions that had been promised in 1972. Although the space shuttle was a remarkable technological achievement as a first-generation reusable launch vehicle, the plans to use it as the only launcher for American payloads proved to be a major policy mistake.

Challenger explosion

The optimism surrounding the space shuttle program was publicly shattered on January 28, 1986, when the *Challenger* orbiter was destroyed in a catastrophic explosion 73 seconds after liftoff. Its seven-person crew perished; among them was schoolteacher Christa McAuliffe, on board as the first private citizen in space. The launch had taken place in unusually cold weather, and a sealing ring within a segment joint of one of the solid rocket boosters failed. The solid rocket broke loose and hit the external tank, rupturing it. The flame from the leaking booster ignited the shuttle's fuel, causing the explosion.

After the accident, the shuttle fleet was grounded until September 1988. A replacement orbiter, *Endeavour,* was built, but, upon the resumption of flights, the shuttle fleet was operated with much greater assurances for the safety of its crew. This limited the flight rate to six to eight missions per year. Both before and after the *Challenger* accident, the space shuttle demonstrated impressive capabilities in space operations, including the repair and redeployment of damaged satellites—most striking being the in-orbit repair of the Hubble Space Telescope in 1993. Beginning in 1998, the space shuttle was used to carry components of the ISS into orbit, along with the crews to assemble those components. It also was used to ferry people and supplies to and from the space station, the role for which it was first conceived.

SCIENCE IN SPACE

The ability to put their instruments into outer space has given scientists the opportunity to acquire new information about the natural universe—information that in many cases would be unobtainable any other way. Space science has added a new dimension to the quest for knowledge, complementing and extending what has been gained from centuries of theoretical speculations and ground-based observations.

Since Yury Gagarin's 1961 flight, space missions involving human crews have carried out a range of significant research, from on-site geological investigations on the Moon to a wide variety of observations and experiments aboard orbiting spacecraft. In particular, the presence in space of humans as experimenters and, in some cases, as experimental subjects has facilitated studies in biomedicine and materials science. Nevertheless, most space science has been carried out by robotic spacecraft in Earth orbit or on missions to various bodies in the solar system. In general, such missions are far less expensive than those involving humans and can carry sophisticated automated instruments to gather a wide variety of relevant data.

In addition to the United States and the Soviet Union, several other countries have the capability to develop and operate scientific spacecraft and thus carry out their own space science missions. They include Japan, Canada, and a number of European countries such as the United Kingdom, France, Italy, and Germany, acting alone and through cooperative organizations involving other European countries. Furthermore, many other countries have become involved in space activities through the participation of their scientists in specific missions. Bilateral or multilateral cooperation among various countries in carrying out space science missions has become the usual way of proceeding.

Categories of space science

Scientific research in space can be divided into five general areas: (1) solar and space physics, including study of the magnetic and electromagnetic fields in space and the various energetic particles also present, with particular attention to their interactions with Earth; (2) exploration of the planets, moons, asteroids, comets, meteoroids, and dust in the solar system; (3) study of the origin, evolution, and current state of the varied objects in the universe beyond the solar system; (4) research on nonliving and living materials, including humans, in the very low gravity levels of the space environment; and (5) study of Earth from space.

Solar and space physics. The first scientific discovery made with instruments orbiting in space was the existence of the Van Allen radiation belts, by Explorer 1 and other spacecraft in 1958. Subsequent space missions carried out by several countries investigated Earth's magnetosphere, the surrounding region of space in which the planet's magnetic field exerts a controlling effect. Of particular interest was—and continues to be—the interaction of the flux of charged particles emitted by the Sun, called the solar wind, with the magnetosphere. Early space science investigations showed, for example, that luminous atmospheric displays known as auroras were the result of this interaction, and scientists came to understand that the magnetosphere is an extremely complex phenomenon.

The focus of inquiry in space physics has since been extended to understanding the characteristics of the Sun, both as an average star and as the primary source of energy for the rest of the solar system, and to exploring space between the Sun and Earth and other planets. The magnetospheres of other planets, particularly Jupiter with its strong magnetic field, also have come under study. Scientists have sought a better understanding of the internal dynamics and overall behaviour of the Sun, the underlying causes of variations in solar activity, and the way in which those variations propagate through space and ultimately affect Earth's magnetosphere and upper atmosphere.

To carry out the investigations required to address these scientific questions, the United States, Europe, the Soviet Union, and Japan have developed a variety of space missions, often in a coordinated fashion. In the United States, early studies of the Sun were undertaken by a series of Orbiting Solar Observatory satellites (launched 1962–75) and the astronaut crews of Skylab in 1973–74. These were followed by the Solar Maximum Mission satellite (launched 1980). ESA developed the Ulysses mission (1990) to explore the Sun's polar regions. In the 1980s, NASA, ESA, and Japan's Institute of Space and Astronautical Science undertook a cooperative venture to develop a comprehensive series of space missions, named the International Solar-Terrestrial Physics Program, that was aimed at full investigation of the Sun-Earth connection. This program was responsible for the U.S. Wind (1994) and Polar (1996) spacecraft, the European SOHO (1995) and Cluster (2000) missions, and the Japanese Geotail satellite (1992).

International Solar-Terrestrial Physics Program

Solar system exploration. From the start of space activity, scientists recognized that spacecraft could gather scientifically valuable data about the various planets, moons, and smaller bodies in the solar system. Both the United States and the U.S.S.R. attempted to send robotic missions to the Moon in the late 1950s. The first four U.S. Pioneer spacecraft, Pioneer 0–3, launched in 1958, were not successful in returning data about the Moon. The fifth mission, Pioneer 4 (1959), was the first U.S. spacecraft to escape Earth's gravitational pull; it flew by the Moon at twice the planned distance but returned some useful data. Three Soviet missions, Luna 1–3, explored the vicinity of the Moon in 1959, confirming that it had no appreciable magnetic field and sending back the first-ever images of its far side. Luna 1 was the first spacecraft to fly past the Moon, beating Pioneer 4 by two months; Luna 2, in making a hard landing on the lunar suface, was the first spacecraft to strike another celestial object. Later Luna spacecraft soft-landed on the Moon, and some gathered soil samples and returned them to Earth.

In the 1960s the United States became the first country to send a spacecraft to the vicinity of other planets—Mariner 2 flew by Venus in December 1962, and Mariner 4 flew past Mars in July 1965. Among significant accomplishments of planetary missions in succeeding decades were the U.S. Viking landings on Mars in 1976 and the Soviet Venera explorations of the atmosphere and surface of Venus from the mid-1960s to the late 1970s. In the years since, the United States has continued an active program of solar system exploration, as did the Soviet Union until its dissolution in 1991. Japan has launched missions to the Moon, Mars, and Halley's Comet. Europe's only independent solar system mission, Giotto, also flew by Halley; in addition, Europe has participated in several U.S. solar system missions.

Three stages of solar system exploration

Early on, scientists planned to conduct solar system exploration in three stages: initial reconnaissance from spacecraft flying by a planet, comet, or asteroid; detailed surveillance from a spacecraft orbiting the object; and on-site research after landing on the object or, in the case of the giant gas planets, by sending a probe into its atmosphere. By the start of the 21st century, all three of those stages had been carried out for the Moon, Venus, Mars, Jupiter, and a near-Earth asteroid. Several Soviet and U.S. robotic spacecraft have landed on Venus and the Moon, and the United States has landed spacecraft on the surface of Mars. A long-term, detailed surveillance of Jupiter and its moons began in 1995 when the U.S. Galileo spacecraft took up orbit around the planet, at the same time releasing a probe into the turbulent Jovian atmosphere. In 2001, the U.S. Near Earth Asteroid Rendezvous (NEAR) spacecraft landed on the asteroid Eros and transmitted information from its surface for more than two weeks. Among the rocky inner planets, only Mercury has been visited just

Galileo and NEAR

once—the U.S. Mariner 10 probe made three flybys of the planet in 1974–75.

As of 2001 the exploration of the other giant gas planets—Saturn, Uranus, and Neptune—remained at the flyby stage. In a series of U.S. missions launched in the 1970s, Pioneer 10 flew by Jupiter, whereas Pioneer 11 and Voyager 1 and 2 flew by both Jupiter and Saturn. Voyager 2 then went on to travel past Uranus and Neptune. The U.S. Cassini spacecraft, launched in 1997, was scheduled to begin long-term surveillance mission in the Saturnian system in 2004, during which its European-built Huygens probe was to descend to the surface of Titan, Saturn's largest moon. Thus, every major body in the solar system except Pluto and its moon, Charon, has been visited at least once by a spacecraft.

What has been learned to date confirms that Earth and the rest of the solar system formed at about the same time from the same cloud of gas and dust surrounding the Sun. The four outer giant gas planets are roughly similar in size and chemical composition, but each has a set of moons that differ widely in their characteristics. The four rocky inner planets had a common origin but have followed very different evolutionary paths and today have very different surfaces, atmospheres, and internal activity. Ongoing comparative study of the evolution of Venus, Mars, and Earth could provide important insights into Earth's future and its continued ability to support life.

Search for extraterrestrial life

The question of whether life has ever existed elsewhere in the solar system continues to intrigue both scientists and the general public. The United States sent two Viking spacecraft to land on the surface of Mars in 1976. Each contained three experiments intended to search for traces of organic material that might indicate the presence of past or present life forms; none of the experiments produced positive results. Twenty years later, a team of scientists studying a meteorite of Martian origin found in Antarctica announced the discovery of possible microscopic fossils resulting from past organic life. Their claim was not universally accepted, but it led to an accelerated program of Martian exploration. A major goal of this program is to return samples of the Martian surface to Earth for laboratory analysis.

The Galileo mission has provided images and other data related to Jupiter's moon Europa that suggest the presence of a liquid water ocean beneath its icy crust. Future missions will seek to confirm the existence of this ocean and search for evidence of organic or biological processes in it.

Exploring the universe. Until the dawn of spaceflight, astronomers were limited in their ability to observe objects beyond the solar system to those portions of the electromagnetic spectrum that could penetrate Earth's atmosphere. These portions included the visible region, parts of the ultraviolet region, and most of the radio-frequency region. The ability to place instruments on a satellite operating above the atmosphere opened the possibility of observing the universe in all regions of the spectrum. Even operating in the visible region, a space-based observatory could avoid the problems caused by atmospheric turbulence and airglow.

Since the 1960s, satellites have been launched by a number of countries to explore cosmic phenomena in the gamma-ray, X-ray, ultraviolet, visible, and infrared regions. More recently, space-based radio astronomy has been pursued. In the last decades of the 20th century, the United States embarked on the development of a series of long-duration orbital facilities collectively called the Great Observatories. They include the Hubble Space Telescope, launched in 1990 for observations in the visible and ultraviolet regions; the Compton Gamma Ray Observatory, launched in 1991; the Chandra X-Ray Observatory, launched in 1999; and the Space Infrared Telescope Facility, to be launched in 2002. Europe and Japan have also been active in space-based astronomy and astrophysics. The results of these space investigations have made major contributions to an understanding of the origin, evolution, and likely future of the universe, galaxies, stars, and planetary systems.

Great Observatories program

Microgravity research. A spacecraft orbiting Earth is essentially in a continual state of free fall. All objects associated with the spacecraft, including any crew and other contents, are accelerating—*i.e.,* falling freely—at the same rate in Earth's gravitational field. As a result, these objects do not "feel" the presence of the Earth's gravity but instead experience a state of weightlessness, or zero gravity. True zero gravity, however, is experienced only at the centre of mass of a freely falling object. With increasing distance from the centre of mass, the influence of gravity increases in directions perpendicular to the object's flight path. These constant but tiny accelerations make necessary the use of the term *microgravity* to describe the space environment. (It is possible to create a similar absence of gravity's effects only briefly on Earth or in an aircraft.) Human activity or operating equipment in a spacecraft causes vibrations that impart additional accelerations and so raise gravity levels, which can make it difficult to carry out highly sensitive experiments under sufficiently low microgravity conditions. Although spacecraft designers cannot totally eliminate gravitational effects, they hope to reduce them in some parts of the International Space Station to one microgravity—one-millionth of Earth's gravity—by isolating those areas from vibrations and other disturbances as much as possible.

The opportunity to carry out experiments in the absence of gravity has interested scientists from the beginning of activities in orbit. In addition to concern about the effects of the weightlessness on humans sent into space, scientists are interested in its effects on the reproductive and developmental cycles of plants and animals other than humans. The overall goal is to use space-based research to add to the general understanding of a wide range of biological processes.

Life-sciences experiments were carried out on the Skylab, Salyut, and Mir space stations and will constitute a significant portion of work aboard the ISS. Such research also has been conducted on space shuttle missions, particularly within the Spacelab facility. In addition, the Soviet Union and United States launched a number of robotic satellites dedicated to life-sciences research. Together these experiments have involved a wide range of nonhuman organisms, from bacteria, plants, and invertebrate animals to fish, birds, frogs, turtles, and mammals such as rats and monkeys. Human crew members also have served as experimental subjects for research on such topics as the functioning of the neurological system and the process of aging. In October 1998, U.S. senator and former Mercury astronaut John Glenn at age 77 returned to space on a shuttle mission dedicated to life-sciences research, which included studies of similarities between the aging process and the body's response to weightlessness.

John Glenn's return to space

The microgravity environment also offers unique conditions for experiments designed to explore the behaviour of materials. Among the areas of inquiry are biotechnology, combustion science, fluid physics, fundamental physics, and materials science. Experiments in the microgravity environment on various materials including metals, alloys, electronic and photonic materials, composites, colloids, glasses and ceramics, and polymers have resulted in a greater understanding of the role of gravity in similar laboratory and manufacturing processes on Earth. Although microgravity research is still largely at the basic level, scientists and engineers hope that additional work—another major focus for the ISS—will lead to practical knowledge of great usefulness to manufacturing processes on Earth.

Observing Earth. Satellites, space stations, and space shuttle missions have provided a new perspective for scientists to collect data about Earth itself. In addition to practical applications (see below), Earth observation from space has made significant contributions to fundamental knowledge. An early and continuing example is the use of satellites to make various geodetic measurements, which has allowed precise determinations of Earth's shape, internal structure, and rotational motion and the tidal and other periodic motions of the oceans. Fields as diverse as archaeology, seismology, and oceanography likewise have benefited from observations and measurements made from orbit.

Scientists have begun to use observations from space as part of comprehensive efforts in fields such as oceanography and ecology to understand and model the causes, processes, and effects of global climate change, including the influence of human activities. The goal is to obtain comprehensive sets of data over meaningful time spans about key physical, chemical, and biological processes that are shaping the planet's future.

SPACE APPLICATIONS

Space visionaries in the early 20th century recognized that putting satellites into orbit could furnish direct and tangible benefits to people on Earth. For example, the English writer Arthur C. Clarke in 1945 described a way in which three satellites above the Equator in a geostationary orbit could relay communications around the globe.

Public-good and commercial applications

Space development, the practical application of the capabilities of spacecraft and of the data collected from space, has evolved in parallel with space exploration. There are two general categories of space applications. One provides benefits that are considered public goods—*i.e.,* that cannot easily be marketed to individual purchasers—and thus are usually provided by governments using public funds. Examples of public-good space applications include meteorology; navigation, position location, and timing; and military and national security uses. The other category of applications provides goods or services that can be sold to purchasers at a profit. These applications are the basis for the commercial development of space by the private sector. Examples of existing commercial space applications include various forms of telecommunications via satellites, remote sensing of Earth's surface, and commercial space transportation. Satellite communications was the first commercial space application, and it has remained the most successful one.

Meteorology. Meteorologists initially thought that satellites would be used primarily to observe cloud patterns and thus provide warnings of impending storms. They did not expect space observations to be central to improved weather forecasting overall. Nevertheless, as the technology of space-based instrumentation became more sophisticated, satellites were called upon to provide three-dimensional profiles of additional variables in the atmosphere including temperature, moisture content, and wind speed. These data have become critical to modern weather forecasting.

Multinational involvement

Meteorological satellites are placed in one of two different kinds of orbit. Satellites in geostationary orbit provide continuous images of cloud patterns over large areas of Earth's surface. From changes in those patterns, meteorologists can deduce wind speeds and locate developing storms. Satellites in lower orbits aligned in a north-south direction, called polar orbits, can obtain more detailed data about changing atmospheric conditions. They also provide repetitive global coverage as Earth rotates beneath their orbit. In the United States, military and civilian agencies each have developed independent polar-orbiting meteorological satellite systems; China, Europe, and the Soviet Union also have deployed their own polar-orbiting satellites. The United States, Europe, the Soviet Union, India, and Japan have orbited geostationary meteorological satellites.

Although the research and development needed to produce meteorological satellites has been carried out by various space agencies, control over satellite operation usually has been handed over to organizations with general responsibility for weather forecasting. In the United States, the National Oceanographic and Atmospheric Administration (NOAA) operates geostationary and polar-orbiting satellites for short- and long-term forecasting, and plans were under way in 2001 to merge that country's civilian and military polar satellite programs under NOAA management. In Europe, an intergovernmental organization called Eumetsat was created in 1986 to operate Europe's meteorological satellites and provide their observations to national weather services.

Positioning, navigation, and timing. In 1957, scientists tracking the first satellite, Sputnik 1, found that they could plot the satellite's orbit very precisely by analyzing the Doppler shift in the frequency of its transmitted signal with respect to a fixed location on Earth. They realized that if this process could be reversed—*i.e.,* if the orbits of several satellites were precisely known—it would be possible to identify one's location on Earth using information from those satellites.

Navstar GPS and GLONASS

This realization, coupled with the need to establish the position at sea of submarines carrying ballistic missiles, led the United States and the Soviet Union each to develop satellite-based navigation systems in the 1960s and early '70s. Those systems, however, did not provide highly accurate information and were unwieldy to use. The two countries then developed second-generation products—the U.S. Navstar Global Positioning System (GPS) and the Soviet Global Navigation Satellite System (GLONASS)—that did much to solve the problems of their predecessors. The original purpose of the systems was the support of military activities, and, at the start of the 21st century, they continued to operate under military control.

The GPS system requires a minimum of 24 satellites, with four satellites distributed in each of six orbits. Deployment of the full complement of satellites was completed in 1994 and included provision for continual replenishment and updating and the maintenance of several spare satellites in orbit. Each satellite carries four atomic clocks accurate to one nanosecond. Because the satellites' orbits are maintained very precisely by ground controllers and the time signals from each satellite are highly accurate, users with a GPS receiver can determine their distance from each of a minimum of four satellites and, from this information, pinpoint their exact location with an accuracy of approximately 18 metres (59 feet) horizontally and 28 metres (92 feet) vertically. The GLONASS system, which became operational in 1996, functions on the same general principles as GPS. A fully deployed system would consist of 24 satellites distributed in three orbits. Because of Russia's economic difficulties, however, GLONASS has not been well maintained and replacement satellite deployment has been slow.

Notwithstanding the military origin of the GPS and GLONASS systems, civilian users have proliferated. They range from wilderness campers, farmers, golfers, and recreational sailors to surveyors, car rental firms, and the world's airlines. The timing information from GPS satellites is also used by the Internet and other computer networks to manage the flow of information. The United States regards GPS as a global utility to be offered free of charge to all users, and it has stated its intent to maintain and upgrade the system into the indefinite future. Concern has been expressed, however, that important worldwide civilian activities such as air-traffic control should not depend on a system controlled by one country's military forces. In response, Europe began in the late 1990s to develop its own navigation satellite system, called Galileo, to be operated under civilian control.

Military and national security uses of space. Those countries and organizations with armed forces deployed abroad were quick to recognize the great usefulness of space-based systems in military operations. The United States, the Soviet Union, the United Kingdom, the North Atlantic Treaty Organization (NATO), and, to a lesser degree, other European countries and China have deployed increasingly sophisticated space systems—including satellites for communications, meteorology, and positioning and navigation—that are dedicated to military uses. In addition, the United States and the Soviet Union have developed satellites to provide early warning of hostile missile launches.

To date, military space systems have served primarily to enhance the effectiveness of ground-, air-, and sea-based military forces. Commanders rely on satellites to communicate with troops on the front lines, and in extreme circumstances national authorities could use them to issue the commands to launch nuclear weapons. Meteorological satellites assist in planning air strikes, and positioning satellites are used to guide weapons to their targets with high accuracy.

Despite the substantial military use of space, no country has deployed a space system capable of attacking a satellite in orbit or of delivering a weapon to a target on Earth. Nevertheless, as more countries acquire military space capabilities and as regional and local conflicts persist around the world, it is not clear whether space will continue to be treated as a weapons-free sanctuary.

Intelligence-gathering satellites

In addition to recognizing the value of space systems in warfare, national leaders in the United States and the Soviet Union realized early on that the ability to gather information about surface-based activities such as weapons development and deployment and troop movements would assist them in planning their own national security activities. As a result, both countries have deployed a variety of space systems for collecting intelligence. They include reconnaissance satellites that provide high-resolution images of Earth's surface in close to real time for use in identifying threatening activities, planning military operations, and monitoring arms-control agreements. Other satellites collect electronic signals such as telephone, radio, and Internet messages and other emissions, which can be used to determine the type of activities that are taking place in a particular location. Most national-security space activity is carried out in a highly secret manner. As the value to national security of such satellite systems has become evident, other countries, such as France, China, India, and Israel, have developed similar capabilities, and still others have begun planning their own systems.

Telstar

Satellite telecommunications. Although some early space experiments explored the use of large orbiting satellites as passive reflectors of signals from point to point on Earth, most work in the late 1950s and early '60s focused on the technology by which a signal sent from the ground would be received by satellite, electronically processed, and relayed to another ground station. American Telephone and Telegraph, recognizing the commercial potential of satellite communications, in 1962 paid NASA to launch its first Telstar satellite. Because that satellite, which operated in a fairly low orbit, was in range of any one receiving antenna for only a few minutes, a large network of such satellites would have been necessary for an operational system. Engineers from the American firm Hughes Aircraft, led by Harold Rosen, developed a design for a satellite that would operate in geostationary orbit. Aided by research support from NASA, the first successful geostationary satellite, Syncom 2, was launched in 1963.

Comsat and Intelsat

The United States also took the lead in creating the organizational framework for communications satellites. Establishment of the Communications Satellite Corporation (Comsat) was authorized in 1962 to operate American communications satellites, and two years later an international agency, the International Telecommunications Satellite Organization (Intelsat), was formed at the proposal of the United States to develop a global network. Comsat, the original manager of Intelsat, decided to base the Intelsat network on geostationary satellites. The first commercial communications satellite, Intelsat 1, also known as Early Bird, was launched in 1965. Intelsat completed its initial global network with the stationing of a satellite over the Indian Ocean in mid-1969, in time to televise the first Moon landing around the world.

The original use of communications satellites was to relay voice, video, and data from one relatively large antenna to a second, distant one, from which the communication then would be distributed over terrestrial networks. This point-to-point application introduced international communications to many new areas of the world, and in the 1970s it also was employed domestically within a number of countries, especially the United States. As undersea fibre-optic cables improved in carrying capacity and signal quality, they became competitive with communications satellites; the latter responded with comparable technological advances that allowed these space-based systems to meet the challenge. A number of companies in the United States and Europe manufacture communications satellites and vie for customers on a global basis.

Direct TV broadcasting

Other space-based communications applications have appeared, the most prominent being the broadcast of signals, primarily television programming, directly to small antennas serving individual households. A similar emerging use is the broadcast of audio programming to small antennas in locations ranging from rural villages in the developing world to individual automobiles. International private satellite networks have emerged as rivals to the government-owned Intelsat, which as of 2001 was itself being transformed into a private-sector organization.

Yet another service that has been devised for satellites is communication with and between mobile users. In 1979 the International Maritime Satellite Organization (Inmarsat) was formed to relay messages to ships at sea. With the growth of personal mobile communications such as cellular telephone services, several attempts were made beginning in the late 1990s to establish satellite-based systems for this purpose. Typically employing constellations of many satellites in low Earth orbit, they experienced difficulty competing with ground-based cellular systems.

Remote sensing. Remote sensing is a term applied to the use of satellites to observe various characteristics of Earth's land and water surfaces in order to obtain information valuable in mapping, mineral exploration, land-use planning, resource management, and other activities. Remote sensing is carried out from orbit with multispectral sensors—*i.e.,* observations are made in several discrete regions of the electromagnetic spectrum that include visible light and usually other wavelengths. From multispectral imagery, analysts are able to derive information on such varied areas of interest as crop condition and type, pollution patterns, and sea conditions.

Landsat and SPOT

Because many applications of remote sensing have a public-good character, a commercial remote-sensing industry has been slow to develop. In addition, the secrecy surrounding intelligence-gathering satellites set stringent limits during the Cold War era on the capabilities that could be offered on a commercial basis. The United States launched the first remote-sensing satellite, NASA's Landsat 1 (originally called Earth Resources Technology Satellite), in 1972. The goals of the Landsat program, which by 1999 had included six successful satellites, were to demonstrate the value of multispectral observation and to prepare the system for transfer to private operators. Despite two decades of attempts at such a transfer, Landsat remained a U.S. government program at the start of the 21st century. In 1986, France launched the first of its SPOT remote-sensing satellites and created a marketing organization, SpotImage, to promote use of its imagery. Both Landsat's and SPOT's multispectral images offered a moderate ground resolution of 10–30 metres (about 30–100 feet). Japan and India also have launched multispectral remote-sensing satellites.

High-resolution remote sensing

In the 1990s, with the end of the Cold War, some of the technology used in reconnaissance satellites was declassified. This allowed several American firms to begin developing high-ground-resolution (less than one metre [3.3 feet]) commercial remote-sensing satellites. The first commercial high-resolution satellite, called Ikonos 1, was launched by the Space Imaging Company in 1999. Among major customers for high-resolution imagery are governments that lack their own reconnaissance satellites. The global availability of imagery previously available only to the leaders of a few countries is troubling to some observers, who express concern that it could lead to increased military activity. Others suggest that this widespread availability will contribute to a more stable world.

Commercial space transportation. The prosperity of the communications satellite business has been accompanied by a willingness of the private sector to pay substantial sums for the launch of its satellites. Initially, most commercial communications satellites went into space on U.S.-government-operated vehicles. When the space shuttle was declared operational in 1982, it became the sole American launch vehicle providing such services. After the 1986 *Challenger* accident, however, the shuttle was prohibited from launching commercial payloads. This created an opportunity for the private sector to employ existing expendable launch vehicles such as the Delta, Atlas, and Titan as commercial launchers. In the years since, a vigorous American commercial space transportation industry has emerged. Whereas the Titan was not a commercial success, the other two vehicles found numerous commercial customers.

Europe followed a different path to commercial space transport. After deciding in the early 1970s to develop the Ariane launcher, it created under French leadership a marketing organization called Arianespace to seek commercial launch contracts for the vehicle. In the mid-1980s, both the

U.S.S.R. and China initiated efforts to attract commercial customers for their launch vehicles. As the industry developed in the 1990s, the United States initiated joint ventures with Russia and Ukraine to market their launchers; Europe also created a similar alliance with Russia.

In 2000 there were 35 commercial launches. In the future, as other countries enter the competition for commercial launch contracts and as new types of launch vehicles are developed, the space business may experience an oversupply of commercial launch capacity. (J.M.Lo.)

BIBLIOGRAPHY

Surface and subsurface exploration. Texts on methodology and discoveries include ARTHUR W. ROSE, HERBERT E. HAWKES, and JOHN S. WEBB, *Geochemistry in Mineral Exploration,* 2nd ed. (1979); ROBERT F. LEGGET and ALLEN W. HATHEWAY, *Geology and Engineering,* 3rd ed. (1988); *Sampling of Soil and Rock* (1971), published by the American Society for Testing and Materials; MARTIN H.P. BOTT, *Interior of the Earth: Its Structure, Constitution, and Evolution,* 2nd ed. (1982); GEORGE D. GARLAND, *Introduction to Geophysics: Mantle, Core, and Crust,* 2nd ed. (1979); ROBERT E. SHERIFF (compiler), *Encyclopedic Dictionary of Exploration Geophysics,* 3rd ed. (1991); ROBERT E. SHERIFF, *Geophysical Methods* (1989); ROBERT E. SHERIFF and L.P. GELDART, *Exploration Seismology,* 2 vol. (1982–83); W.M. TELFORD, L.P. GELDART, and ROBERT E. SHERIFF, *Applied Geophysics,* 2nd ed. (1980); CHUJI TSUBOI, *Gravity* (1983; originally published in Japanese, 1979); W.D. PARKINSON, *Introduction to Geomagnetism* (1983); J.A. JACOBS, *A Textbook on Geonomy* (1974); and FLOYD F. SABINS, JR., *Remote Sensing: Principles and Interpretation,* 2nd ed. (1987). (R.E.Sh.)

Undersea exploration. The classical tools of the oceanographer are described by H.U. SVERDRUP, MARTIN W. JOHNSON, and RICHARD H. FLEMING, *The Oceans* (1942, reissued 1970). More recent texts on oceanography and marine biology include M. GRANT GROSS, *Oceanography, a View of the Earth,* 5th ed. (1990); GEORGE L. PICKARD and WILLIAM J. EMERY, *Descriptive Physical Oceanography,* 5th enlarged ed. (1990); JAMES P. KENNETT, *Marine Geology* (1982); BRUCE A. WARREN and CARL WUNSCH (eds.), *Evolution of Physical Oceanography* (1981); JAMES L. SUMICH, *An Introduction to the Biology of Marine Life,* 5th ed. (1992); M.N. HILL *et al.* (eds.), *The Sea* (1962–); and J.P. RILEY and G. SKIRROW (eds.), *Chemical Oceanography,* 2nd ed., 8 vol. (1975–83).

Exploration and discovery are chronicled in RACHEL L. CARSON, *The Sea Around Us,* special ed. (1989); JAMES DUGAN and RICHARD VAHAN (eds.), *Men Under Water* (1965); and HARRIS B. STEWART, JR., *Deep Challenge* (1966). Also informative is the documentary of famous explorers, such as the books by WILLIAM BEEBE, *Half Mile Down* (1934, reissued 1951); JACQUES PICCARD and ROBERT S. DIETZ, *Seven Miles Down* (1961); GEORGE STEPHEN RITCHIE, *Challenger: The Life of a Survey Ship* (1957); HELEN RAITT, *Exploring the Deep Pacific,* 2nd ed. (1964); WILLARD BASCOM, *A Hole in the Bottom of the Sea: The Story of the Mohole Project* (1961); FRANCIS P. SHEPARD, *The Earth Beneath the Sea,* rev. ed. (1967); and DANIEL BEHRMAN, *The New World of the Oceans* (1969). Information on exploration and survey vessels in operation today may be found in an annually updated loose-leaf catalog, *Oceanography Vessels of the World* (1961–), published by the U.S. Naval Oceanographic Office. STEWART B. NELSON, *Oceanographic Ships, Fore and Aft* (1971, reprinted 1983), provides a complete history of U.S. vessels up to the time of publication. See also JACQUES COUSTEAU and ALEXIS SIVERINE, *Jacques Cousteau's Calypso* (1983; originally published in French, 1978); R. FRANK BUSBY, *Manned Submersibles* (1976), on their design and operation; and KENNETH J. HSÜ, *Challenger at Sea,* trans. from German (1992).

Modern undersea explorers are strongly dependent on innovative engineering and technology, such as that treated in *Jane's Ocean Technology, 1979–80* (1979); JOHN J. MYERS (ed.), *Handbook of Ocean and Underwater Engineering* (1969); JOHN F. BRAHTZ (ed.), *Ocean Engineering* (1968); and ROBERT L. WIEGEL, *Oceanographical Engineering* (1964). See also JOHN BRACKETT HERSEY (ed.), *Deep-Sea Photography* (1967); J.F.R. GOWER (ed.), *Oceanography from Space* (1981); G.A. MAUL, *Introduction to Satellite Oceanography* (1985); and F. DOBSON, L. HASSE, and R. DAVIS (eds.), *Air-Sea Interaction: Instruments and Methods* (1980). (A.B.R./D.J.Ba./Ed.)

Space exploration. Broad coverage of space activities can be found in MICHAEL RYCROFT (ed.), *The Cambridge Encyclopedia of Space* (1990; originally published in French, 1987). An overall history of space exploration is WILLIAM E. BURROWS, *This New Ocean: The Story of the First Space Age* (1998). WALTER A. MCDOUGALL, *The Heavens and the Earth: A Political History of the Space Age* (1985, reissued 1997), traces the U.S.-Soviet rivalry that led to the space race and comments on its impact on the two countries' societies. Earlier historical discussions include WILLY LEY, *Rockets, Missiles, and Men in Space,* newly rev. and expanded ed. (1968); and WERNHER VON BRAUN, FREDERICK I. ORDWAY III, and DAVID DOOLING, *Space Travel: A History,* 4th ed. (1985). FRANK H. WINTER, *Rockets into Space* (1990), provides an account of the development of rocketry. Speculative discussions of the promises of space exploration include ARTHUR C. CLARKE (compiler and ed.), *The Coming of the Space Age: Famous Accounts of Man's Probing of the Universe* (1967, reissued 1970); HARRY L. SHIPMAN, *Humans in Space: 21st Century Frontiers* (1989); CARL SAGAN, *Pale Blue Dot: A Vision of the Human Future in Space* (1994, reissued 1997); and ROBERT ZUBRIN and RICHARD WAGNER, *The Case for Mars: The Plan to Settle the Red Planet and Why We Must* (1996).

Many early American astronauts have written of their experiences. The best of these works is MICHAEL COLLINS, *Carrying the Fire: An Astronaut's Journeys* (1974, reissued 2001). An account of the Apollo program that is focused on astronauts is ANDREW CHAIKIN, *A Man on the Moon: The Voyages of the Apollo Astronauts* (1994, reissued in 3 vol., 1999); this book served as the basis for a video series, *From the Earth to the Moon* (1998), produced by TOM HANKS. A failed Apollo mission is the subject of the theatrical film *Apollo 13* (1995). The best account of Apollo from the perspective of its managers and engineers is CHARLES MURRAY and CATHERINE BLY COX, *Apollo: The Race to the Moon* (1989). A noted author offers his impressions of Apollo in NORMAN MAILER, *Of a Fire on the Moon* (1970, reissued 1985; also published as *A Fire on the Moon,* 1970). TOM WOLFE, *The Right Stuff,* new ed. (1983, reissued 1997), provides an account of the early days of U.S. human spaceflight. JOHN M. LOGSDON, *The Decision to Go to the Moon: Project Apollo and the National Interest* (1970, reissued 1976), traces the political underpinnings of the Apollo program. An extensive account of the Soviet space program during the race to the Moon is ASIF A. SIDDIQI, *Challenge to Apollo: The Soviet Union and the Space Race, 1945–1974* (2000).

The origins of U.S. post-Apollo spaceflight programs are discussed in T.A. HEPPENHEIMER, *The Space Shuttle Decision: NASA's Search for a Reusable Space Vehicle* (1999); and HOWARD E. MCCURDY, *The Space Station Decision: Incremental Politics and Technological Choice* (1990). Events leading to the 1986 *Challenger* accident are detailed in JOSEPH J. TRENTO, *Prescription for Disaster* (1987). A selective view of U.S.-Russian cooperation in human spaceflight is found in BRYAN BURROUGH, *Dragonfly: NASA and the Crisis Aboard Mir* (1998, reissued 2000).

Available in addition to the works cited above are published studies, sponsored by the NASA History Program, of almost every one of the agency's space programs. Original documents tracing the history of the U.S. space program are reprinted in JOHN M. LOGSDON *et al.* (eds.), *Exploring the Unknown: Selected Documents in the History of the U.S. Civil Space Program* (1995–).

A comprehensive discussion of European space activities up to 1987 is provided in J. KRIGE, A. RUSSO, and L. SEBESTA, *A History of the European Space Agency 1958–1987,* 2 vol. (2000). ROGER M. BONNET and VITTORIO MANNO, *International Cooperation in Space: The Example of the European Space Agency* (1994), elaborates on international space activities from a European perspective. Other national space efforts are described in BRIAN HARVEY, *The Chinese Space Programme: From Conception to Future Capabilities* (1998), *The Japanese and Indian Space Programmes: Two Roads into Space* (2000), and *Russia in Space: The Failed Frontier?* (2001).

Discussions of various space science efforts include HOMER E. NEWELL, *Beyond the Atmosphere: Early Years of Space Science* (1980); BRUCE MURRAY, *Journey into Space: The First Three Decades of Space Exploration* (1989); and WILLIAM E. BURROWS, *Exploring Space: Voyages in the Solar System and Beyond* (1990); the last two works deal with U.S. missions to explore the solar system. Also pertinent is ROBERT W. SMITH, *The Space Telescope: A Study of NASA, Science, Technology, and Politics* (1989, reissued 1993).

The origins of reconnaissance satellite programs are covered in DWAYNE A. DAY, JOHN M. LOGSDON, and BRIAN LATELL (eds.), *Eye in the Sky: The Story of the Corona Spy Satellites* (1998). Subsequent spy satellite programs are discussed in JEFFREY T. RICHELSON, *America's Secret Eyes in Space: The U.S. Keyhole Spy Satellite Program* (1990); and WILLIAM E. BURROWS, *Deep Black: Space Espionage and National Security* (1986, reissued 1988). Early debates over the military use of space are described in PAUL B. STARES, *The Militarization of Space: U.S. Policy, 1945–1984* (1985); and more recent debates on this issue are summarized in PETER L. HAYS *et al., Spacepower for a New Millennium: Space and U.S. National Security* (2000).

HEATHER E. HUDSON, *Communication Satellites: Their Development and Impact* (1990), gives a synopsis of progress in communications satellites. Controversies surrounding the development of Earth observation satellites are followed in PAMELA E. MACK, *Viewing the Earth: The Social Construction of the Landsat Satellite System* (1990). (J.M.Lo.)

Family and Kinship

Family and kinship are among the most important aspects of human society. They play a central part in the social organization of peoples throughout the world. Nevertheless, both the organization of the family unit and the structure of kinship relations vary from society to society and through time.

The study of family and kinship falls particularly within the scope of three academic disciplines: sociology, anthropology, and social history. Sociological studies tend to concentrate on the form and organization of the modern family and on the social problems that surround family life. Anthropological studies also frequently concern the family, but their emphasis is on the variety of family organizations and on cross-cultural comparisons. Anthropologists are also interested in kinship ties and obligations beyond the immediate family, since these are often extremely important in small-scale nonindustrialized societies, on which anthropological studies tend to concentrate. Social and economic historians, particularly those who practice in the subdiscipline known as family history, utilize the perspectives of a variety of disciplines, including psychology and economics, as well as sociology and anthropology.

This article discusses the history and structure of the family, with emphasis on the Western family, although comparisons are made with family and kinship structures of other cultures. Theoretical approaches to the study of family and kinship are also discussed, as are the types of kinship terminologies, descent systems, and marriage systems found in cultures throughout the world. For discussions of the family in legal systems, particularly those of the West, see the articles FAMILY LAW and INHERITANCE AND SUCCESSION. Many of the stages of life discussed here in their familial contexts are also treated in BEHAVIOUR, THE DEVELOPMENT OF HUMAN. For a discussion of reproductive and emotional aspects of the family, see SEX AND SEXUALITY. For a discussion of family planning, see BIRTH CONTROL.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 513 and 522, and the *Index*.

The article is divided into the following sections:

Family history

Western society since medieval times has been characterized by a great diversity of family organization. This diversity has had several often interrelated aspects, including geographic region, occupation, social class, and whether the family in question was rural or urban. Historically, there were considerable differences between different regions of Europe and even between different areas in the same country. Aristocrats and commoners established family customs and social institutions peculiar to their respective social classes, as did merchants and peasants. In addition, ethnic and religious minorities frequently established their own unique patterns of family life in accord with traditions and moral values that were often at variance with the rest of society.

APPROACHES TO FAMILY HISTORY

The task of interpreting and explaining the diversity of family organization in historical times has fallen to the field of family history. Family history emerged in the 1960s as a major focus of interest within the broader field of economic and social history. Because family history draws on the methods and theories of several social sciences, it has developed its own diversity of approaches to the understanding of its subject. Three major areas of interest are relevant here; the British family historian Michael Anderson labeled them the sentiments approach, the demographic approach, and the household economics approach.

The sentiments approach. The sentiments approach emerged from the interest of many scholars in the emotional ties between family members. Proponents of this approach try to understand the character of the family as it has changed through time. Topics of interest include the nature of conjugal and parent–child relationships, courtship practices, attitudes toward sex, and the relative importance of privacy and individualism in the family context.

Most proponents of the sentiments approach argue that changes in family behaviour are the result of changes in other aspects of culture. The wider society, as it undergoes change, creates cultural values affecting the family. Religious and philosophical ideas, for example, may play a part in shaping attitudes toward individualism or social equality. Even laissez-faire capitalism, according to some writers, played a part in the development of trends toward sexual freedom. General social trends and social revolutions, such as the Reformation and the Industrial Revolution, ultimately affect the behaviour and outlook of ordinary people in their family relations. The great difficulty in pursuing this approach, however, is that documentary sources are not always helpful in providing the kind of information these arguments require. Relating broad social changes to changes in sentiment is often done only through educated speculation.

Studying parish registers

The demographic approach. This approach favours more limited objectives, and its methods are akin to those of the natural sciences. Rather than literary sources, which are more frequently used by those who follow the sentiments or household economics approach, demographers concentrate on available data on households, baptisms, marriages, and burials. Parish registers in many European countries provide a large amount of source material of this kind, often going back to the 16th century. Historical demographers can use these data to build up a picture of family life at any given place and time—a particular village, for example, or a specific occupational group or social class.

The main problem of this approach is that the available data may be too incomplete or inaccurate for making useful generalizations. The original documentation may have been made for purposes of taxation or military recruitment; certainly it was not made with the interests of historians in mind. As a result, the technique of the historical demographer depends on "family reconstitution." If a range of sources is present, the reconstruction of household size and composition, marriage patterns, and a number of other features of family organization is possible, but it is still problematic.

The family as an economic unit

The household economics approach. This approach is the broadest of the three. Its proponents try to take into greater account the methods and theories of other social sciences, especially anthropology (also an interest of some proponents of the sentiments approach), with its comparative study of social relationships and the life cycle of the family. In essence, the household economics approach is concerned with the ways in which the members of the family or household interact with one another in an economic context. Objects of study include relations between husband and wife, master–servant relations (if servants are present in the household), and parental control over the labour of children. Inheritance of property, succession to titles of nobility, and rights to membership in wider family and kin groupings are also areas of special interest, as is the place of the household within the economic life of the community.

While the three approaches outlined above are in many ways complementary, the household economics approach is of the greatest utility for the study of family history from the earliest written records. The sentiments and demographic approaches developed primarily with reference to the history of the Western family since the 16th century, because data useful to these approaches are extremely difficult to obtain from before that time. The household economics approach, therefore, is the main one taken in this article to outline the history of the family in western Europe.

HISTORY OF THE FAMILY IN WESTERN EUROPE

The family in classical antiquity. Within the ancient Greek city-states, four levels of kin grouping could be distinguished: the phratry (*phratra*), the aristocratic clan (*genos*), the kindred (*anchisteis*), and the household (*oikos*). The phratries were large tribal subdivisions consisting of households or families claiming a common kin relation. At the apex of each phratry were aristocratic clans that had certain hereditary rights, such as the right to hold priestly offices. Both the phratries and the clans were recruited patrilineally (through the father).

The kindreds, on the other hand, consisted of a set of relatives from each side of the family, extending at least to the second cousins of any given individual. A person's kindred was important, as this was the unit within which inheritance could be claimed, but the significance of the kindred varied from city-state to city-state and through history. For example, during the 5th and 4th centuries BC such ties of kinship were less important in urban Athens than elsewhere, because of increasing urbanization and the large number of noncitizens in the population. Since in theory only citizens could contract legal marriages and produce legitimate children, the large number of common-law marriages between citizens and noncitizens (and attempts to pass off the offspring as legitimate) led, in classical Athens, to the dissolution of kindred ties that had previously been close.

The organization of the household also varied from place to place and through history. From the time of the Homeric poems (before 700 BC) it seems that the primary unit of residence and domestic economy was the nuclear family (husband, wife, and children). However, a belief in earlier extended-family organization is implied in the *Iliad* and *Odyssey* by the archaic kinship terms employed for relatives of a king and by descriptions of such large extended families as those of Zeus on Mount Olympus and of Priam, king of Troy. In the Homeric world the household was the seat of a person's prestige and a political unit through which alliances were made and struggles for dominance played out.

The rise of the household

As political institutions developed and replaced the household as the seat of power and intrigue, the household became a more private place. The development of the city-states, and especially of Athens, tended to separate public and private life in a way that had not been common before. Public life in the Athenian democracy was characterized by equality among fellow citizens but, at the same time, by an impersonal and competitive attitude toward one another as well as toward noncitizens. Private life (within the household) was characterized by hierarchy, intimacy, and support. The household was hierarchical in that it was headed by a man who held authority over his wife and children and his servants. The ancient Greek household was the key unit of economic relations, too—as is implied in the modern English word economics, from the Greek *oikonomia,* meaning "household management."

The English word family is derived from the Latin word *familia,* meaning "household," and ultimately from the Latin *famulus,* "servant." As it was among the Greeks, the household among the Romans was a significant economic unit within a system of wider kinship ties. The patrilineal kin group or clan among the Romans was termed the gens (plural, gentes). For the aristocratic Roman, this level was more important than was the phratry or clan among the Greek aristocracy, and together the gens and the household were primary focal points of life in ancient Roman society.

The Roman gens comprised a group of kinsmen who bore a common gens name, inherited in the male line by both males and females. In the earliest times, each gens had one or more recognized chiefs. The gentes themselves were grouped into larger units, each known as a curia (roughly equivalent to the Greek phratry), and, above that level, into "tribes," though the significance of these units declined after the formation of the Roman republic in 509 BC. According to custom, each gens shared common property, which included a burial ground and other lands. Property was inherited within the gens, and gens members could call upon one another for help in defending their individual property and for the redress of injuries. It was also forbidden for people to marry within their own gens. Although in theory the gens was patrilineal, under certain conditions and with the consent of the gens as a whole, outsiders, such as kinsmen related other than through the male line, could be adopted into the gens.

Importance of the agnatic (male) line

Whereas the gens as a corporate group was important, especially in the early days of the Roman state (later being important only for aristocrats), descent in the male line was recognized as a significant legal principle by both commoners and aristocrats throughout Roman history. The Romans distinguished two kinds of blood relatives: agnates and cognates. An agnate (Latin *agnatus*) was a relative related through the male line, including one's father's father, father's brothers and sisters, father, brothers, sisters, children (of a man but not of a woman), brothers' children, sons' children, and some more distant relatives. A cognate (*cognatus*) was any other blood relative. Even Romans who did not belong to an aristocratic gens distinguished their cognates from their agnates. Agnates were reckoned to be "closer," because private property was inherited by them and marriage was forbidden between them.

The medieval family. The medieval European family was a product of diverse origins and influences. The family and kinship customs of the Germanic tribesmen, the legal system inherited to a large extent from the Romans, the ideals of Christianity and dominance of the church, and the emergence of feudalism all played a part in shaping the European family of the medieval and postmedieval periods.

The countries of the eastern Mediterranean retained the ideals of patrilineal descent to a greater extent and for much longer than those to the west, where an emphasis on bilateral descent (equal reckoning of kinship from both sides of the family) replaced the more formal structure of the Roman gentes. In northern Europe the Germanic and probably also the Celtic tribes were organized largely on a bilateral basis. Membership in the kin group could be achieved through either one's father or one's mother. The ancient clans of the Irish and Scots, for example, were actually almost tribes, with marriage permitted inside or outside the clan and membership often being determined by residence rather than strictly by descent. Marriage, in fact, was permitted to close kin among the pre-Christian Germanic tribesmen, a practice that no doubt reinforced family ties. Among the Germanic tribesmen monogamy was the norm, but concubines were also permitted.

With the spread of Christianity, close kin marriage came to be forbidden, under penalty of slavery, though concubinage continued for several centuries. (Though they were not allowed to marry, even priests openly took concubines and raised families.) From the 4th century, close kin forbidden as spouses included first cousins, certain relatives through marriage, and even "fictive" kin such as godchildren. In the 11th century the prohibitions were extended yet further, though they were contracted again in the 13th century. Divorce, which had been permitted by the Greeks, the Romans, and the tribesmen of northern Europe, also came to be forbidden or, initially, at least made extremely difficult.

Breaking up the kin group

Of all these changes, the prohibition on close kin marriage was perhaps the most significant. Both St. Augustine and St. Thomas Aquinas argued that marriage outside the group increased kinship ties between communities, thus breaking down the undesirable solidarity of close-knit groups. The church, in a sense, broke the authority of the extended family and kin group. Marriage outside the group also tended to disperse rather than concentrate inherited property and thereby lessened the power and influence of kin-based groups. It is difficult to determine whether these reforms were introduced with the intention of decreasing the authority of families and close-knit communities in preference to the authority of the church or whether this consequence was merely coincidental. Certainly they were not specifically justified on scriptural grounds. Whatever their reason, the changes in marriage prohibitions dramatically altered the nature of the family in much of western Europe, changing it from an extended and essentially self-contained unit with authority over its own members to a smaller, nuclear unit ceding much of its former authority to the church.

Emergence of the nuclear family

Another long-term effect of the decrease in the solidarity of larger kin groups was the increase in importance of the conjugal bond, which in the Middle Ages replaced the large kin group as the focus of familial ties. Here, too, the church was a major force in change, for example, in requiring that those who were to be married consent to the union. Thus, marriage arrangements could not rest solely in the hands of household heads, although marriage was certainly much more a matter for the family as a whole than it is in modern Europe. A gradual change toward the modern family took place in the late Middle Ages, with women and children being recognized as having rights of their own, independent of the head of the household, and with the emergence of the nuclear family or small, core kin group as the basic unit of both production and social life. (The Black Death, which killed about a third of the inhabitants of Europe between 1347 and 1351, may also have been a factor, but recent research suggests that family practices were much less affected by the Black Death than were other social institutions. In fact, changes toward nuclear family organization had taken place in northern Europe long before the 14th century.)

Accompanying these changes was a change in custom regarding the transfer of property at marriage. The Germanic peoples had a custom of paying bridewealth, livestock and other goods transferred at marriage from the bridegroom and his family to the family of the bride. This practice occurred in much of Europe through the Middle Ages, but toward the end of this period there was a marked trend instead toward dowry, payment from the bride's family to the groom for the upkeep of his new spouse. In some cases both kinds of payment existed simultaneously. The reasons for the shift in emphasis are a matter of debate among historians and are no doubt related in complex ways to other social trends in the organization of the domestic economy and of society in general.

The family since 1500. Changes in the family that were set in motion in medieval times continued through the Renaissance and Reformation. The conjugal bond remained important. The question of how much affection existed between husbands and wives has been the subject of much speculation, especially among those social historians who favour the sentiments approach. The British historian Lawrence Stone suggested that the degree of affection in the 16th and early 17th centuries was very limited, but other writers have been critical of this view. The upper-class domestic unit, in fact, seems to have been a close-knit one, but adequate data on the lower classes is extremely hard to find.

Marriage between cousins

One consequence of the Reformation was a reduction in the strictness of rules regarding the marriage of kin, at least in most Protestant countries. The case of England's King Henry VIII is a famous example. He legalized marriage to first cousins so that he could marry Catherine Howard, who was in fact his "first cousin" by marriage, being the true first cousin (in the modern sense of a blood relative) to one of his previous wives, Anne Boleyn. Martin Luther, on the other hand, argued against close kin marriage, not because he believed it was against God's will, but because he was afraid that people would marry in order to keep property within the family, rather than for love.

Sociologists used to assume that before the Industrial Revolution the most common type of family organization in European society was that of the large extended family, with as many as three generations sharing a common household. However, detailed historical studies made in the 1970s and '80s from documentary sources have revealed that the nuclear family was the primary unit over much of northwestern Europe even before industrialization. In fact, Anderson's study of the English town of Preston suggests that the Industrial Revolution, and its consequent urbanization and urban employment, increased the importance of relatives beyond the nuclear family, since the changes taking place enabled individuals to receive help from their kinsmen in times of hardship. In general, in 19th-century Europe, North America, and other parts of the industrialized world, family organization was based on a wider range of social interaction than was usual by the late 20th century. Since marriage was relatively late and life expectancy relatively short, children often remained at home throughout the lifetimes of their parents. Families were large, and older children took part in raising and teaching their younger brothers and sisters. Only in the 20th century has it become common for parents to be left alone in the household in middle age.

In the United States there was an increased tendency toward earlier marriage for both sexes between 1900 and 1960. Since about 1960, however, this trend has reversed. Similar trends have occurred, though generally later, in the United Kingdom and other Western countries. A number of factors seem to be involved in this trend toward later marriage. One is greater economic security. Another is the greater availability of birth control devices, and especially the birth control pill, since the 1960s. This has enabled couples to engage in sexual intercourse with less fear of having unwanted children. Whether the practice of birth control has been a cause or a result of greater permissiveness is not easy to determine, but certainly greater sexual freedom without the social stigma formerly attached to premarital intercourse is another factor in the trend toward later marriage.

Increase in age of marriage

There have long been differences between social classes in the customary age at which people marry. Generally, since 1900, those from lower socioeconomic backgrounds have married earlier. The reason usually given for this fact is that people prefer to marry after their education has ended and they have secured suitable employment. Middle- and upper-class people tend to be engaged in full-time education for a longer time, and career aspirations may outweigh the importance of marriage. In some lower-class households, where marriage itself may be a person's greatest aspiration, this is less likely to be the case. As women achieve educational standards equal to those of men and enter the job market on the same footing, they tend to regard marriage as something to be delayed until "the right time," just as men do. The increase in educational opportunities for people of both sexes since World War II may be an additional factor in the increase in the age of marriage in all social classes within Western societies.

The modern family

There is no precise distinction between modern and premodern families. Some authorities consider the modern family to be identical with the Western family since the Industrial Revolution. Others take the notion of modern in a more literal sense, meaning the family as it is today in any part of the world. Certainly there is no reason to exclude the entire Third World and equate modern only with Western society.

Yet too precise a definition is a hindrance rather than a help. In different contexts the word modern means subtly different things. Its definition depends on the contrasts and comparisons to be made. For the purposes of the following discussion, the "modern family" may be taken generally to refer to family organization as it has existed since the early 20th century. Special concern is directed, mainly but by no means exclusively, toward family organization in Western societies. Even so, cultural, subcultural, religious, and class differences create a variety of forms within the modern family, and this variety cannot be overestimated.

THE FAMILY CYCLE

Courtship and mate selection. In most societies adolescence is marked by social as well as biologic changes. Such social changes often include a new, more formal attitude toward parents and, more particularly, toward members of the opposite sex. In many societies the custom is for young people in the early stages of adolescence to spend time in same-sex groups. Boys may attend sporting events, and girls, overnight parties. Eventually the same-sex groups come into contact with similar groups of the opposite sex at social occasions. In some societies young people go out with mixed-sex groups before they begin dating, but in other countries the practice of dating emerges from the same-sex group phase, as individual young people meet and pair off.

Social aspects of sexual awakening

Often dating is followed by a stage of "steady dating," in which a couple agree to date only each other, and their exclusive dating relationship may become recognized by their peer group and others, including their respective parents. If the couple gets along well together, the phase sometimes defined as courtship may begin.

In its strictest sense courtship usually refers to an activity, such as dating, with intent to marry. Prior to the 20th century, and still today in some societies, courtship was practiced without dating in the modern sense. A young man might visit his intended bride in her parental home and bring gifts, discuss common interests, and perhaps go for walks. To nontraditional people these customs, if continued to courtship, may seem ridiculously old-fashioned, and it is frequently difficult in the modern era to make meaningful distinctions between "going out," dating, and courting. The distinctions are further blurred by modern couples' living together, either before marriage or with no intention of ever marrying.

More typically, though, the family begins with marriage, which grows out of courtship and is preceded by what sociologists call mate selection. Mate selection may be defined simply as the process, often unconscious, of choosing a mate. Usually the mate is the person's intended spouse, and the traditional definition of the term has this implication. Individuals often claim that their match is made on the basis of love, but statistical studies show that other factors are involved. For example, people usually marry within their social class and often to those of the same ethnic group or religion. This practice of like marrying like is known as homogamy. Mate selection is also frequently related to economic factors. For instance, before she will consent to marry him, a woman may want her intended husband to show that he is capable of supporting her.

Homogamy: Cupid and statistics

Betrothal stage. Betrothal may be defined as the recognition by both parties of their intent to marry. It implies a mutual obligation to marry, and it sometimes involves a formal contract between the respective families of the couple. Often an engagement ring, marking this obligation, is given by the man to his intended bride. The ring not only signifies their intent but also publicly identifies the couple—particularly the woman who wears the ring—as intended spouses, warning off other potential suitors. Traditionally, a woman wears her engagement ring with great pride. Yet perhaps because of the inequality in the signification of the event (the fact that, usually, only the woman wears a ring), the custom of giving and receiving such a ring is less common than it used to be.

Just as it has become difficult to draw the boundary between dating and courtship, so too it is sometimes difficult to draw a rigid distinction between courtship and betrothal. A couple may live together for a number of years before deciding to marry, and the betrothal stage and even the marriage itself may thus be reduced from its traditional significance. This is especially true in Western societies, particularly for young middle-class couples. The decline in importance of betrothal is linked to the greater degree of sexual freedom permitted in many societies today. Previously, betrothal often allowed a couple to engage in sexual activities not permitted between those who were merely dating or courting. Even where sexual intercourse was prohibited before marriage, other forms of sexual license or nonsexual intimacy were permitted. In parts of Scotland and Scandinavia, for example, engaged couples were allowed to sleep in the same bed but were sewn up in different sleeping bags, a custom known as "bundling."

Decline of betrothal

Marriage stage. Marriage is frequently regarded as the most important event in a person's life. It is also a significant event in the family cycle, both for the families of the bride and groom and for the new family formed by the union itself. For young people, marriage often marks a break with the authority of their parents (who, together with brothers and sisters, are known as the "family of orientation"), and it is simultaneously a key step in the formation of a new family (the "family of procreation"). The couple must learn to live together and work out their own changing social roles within the context of the marital unit. Thus, though marriage often represents a break with parental authority, it also involves a sacrifice of independence. Husbands and wives must compromise with each other and learn to make decisions, not alone, as they had done when single, but together. Perhaps not surprisingly, those who marry for the first time relatively late in life often retain more individualism than those who marry early. Such people may maintain a greater degree of social independence and at the same time develop successful marital and family relationships.

In the early period of marriage, husbands and wives tend to spend a great deal of time with each other. It is traditional in some societies for a husband or wife to have a "night out," perhaps once a week, with other people of the same sex and without the company of the spouse. The rest of the week, except for working hours, they generally spend together. This time together, especially if begun at a relatively early age, leads to unconscious learning and acceptance of each other's habits and idiosyncracies.

Marriage is usually marked by a sharp increase in sexual activity. Sexual access is obviously easier for a married couple living together than for most single people, who tend to live apart from their partners. For many people, for reasons of religion or morality, sexual intercourse is permissible only after marriage. For others, marriage simply makes sexual activity more convenient.

Social status of marriage

Marriage also marks a change in social relations with others. Whereas before marriage people interact in most social spheres as individuals, after marriage there is a tendency for the married couple to be the primary unit in social activities. Invitations to parties, for example, are made to a couple, not to one partner only (though this may also be true of engaged or courting couples). The couple may come to be seen as an inseparable unit, thus marking the marital unit as a "corporate" entity. Recognition of this status also paves the way for the recognition of the role of the marital unit in its ultimate capacity as a family of procreation.

Children. For most people, children are the key to a happy family life. They are also integral to the definition of family life, and many sociologists regard the raising of children as the primary function of the family.

Childbirth, of course, is a biologic fact, but it is equally a social phenomenon. It requires a readjustment of roles as people become not only husband and wife to each other but also parents to the newborn child. The birth of subsequent children can have equally considerable restructuring effects on the family, including a change in roles for the older children. Relations between the new parents and their own parents are often heightened as the latter take on their new roles as grandparents. In many societies, grandparents are indulgent toward their grandchildren, while the parents themselves may be required by convention, and some would say by necessity, to be stricter.

The specific social adjustments people must make after the birth of a child differ according to individual circumstances and cultural attitude. No matter what pattern is followed, however, ideally the children are brought up to become productive members of their society. To achieve this, they must acquire education within the family and, later, within the institutions of the society as a whole.

Child rearing and socialization. The American sociologist Talcott Parsons believed that the two most important functions of the modern family are the primary socialization of children and the stabilization of adult personalities through marriage and the raising of children. His own concern was particularly with the middle-class American family, but these important aspects of family life are also applicable much more widely. In the present context it is worthwhile to look especially at primary socialization.

Primary socialization refers to the training of children during their earliest years, whereas secondary socialization refers to later influences on the development of the child's personality and learning activities, such as his involvement with teachers and with other children at school. Primary socialization is in most societies carried out essentially within the family as part of child rearing. In the modern family, parents take responsibility for raising and teaching their children such basic things as language and correct behaviour. Toilet training, teaching children how to eat correctly, and encouraging children to get along with others are all aspects of child rearing. However, it is not only these more mundane aspects of behaviour that children learn. Children are also implicitly encouraged to develop the values of the parents and of the society in which they live. In American society, which was Parsons' main concern, these values include independence, motivation for achievement, and competition. In other societies, different values, such as cooperation and egalitarianism, may be stressed. Yet the principle behind primary socialization in different societies is the same: the development of social values must be achieved in an environment of love and security, as is found in the ideal family anywhere in the world.

The child as scapegoat

Few families are ideal, however. Studies of the families of emotionally disturbed children have shown that unsatisfactory relationships between husbands and wives can have detrimental effects on children. Sometimes a child is used as a scapegoat. The parents blame or even physically abuse the child in order to cover up their own difficulties. In such a case, the child often fails to develop the values the parents wish to instill in him, developing instead antisocial habits leading to deviant behaviour in later life. Indeed, the cycle may be repeated if such a person in time marries, has a family of his own, and treats his children in the same way. Nonetheless, there is no reason to suppose that all children of unsatisfactory marriages are treated in such a way or fail to overcome the difficulties they have as children.

Some social scientists have even suggested that the isolated nuclear family, as it exists in Western industrialized societies, is to blame for the social ills found in those societies. They claim that in the past more support was offered from the wider kin network and from the community as a whole—as is still the case in less-developed parts of the world. The British psychiatrists R.D. Laing and David Cooper suggested that the modern family is dysfunctional in that, by its very nature, it forces upon children an undue emphasis on obedience to authority. These negative viewpoints aside, most experts as well as most parents agree that the primary socialization process in the modern family offers benefits both to the child and to the parents.

Middle age. As the children grow up, the family undergoes certain changes. Parental roles shift from child rearing and socialization to financial maintenance coupled with keeping a watchful eye on the children's school and social activities and preparing them for adulthood.

Middle age is generally considered to be the stage in the life cycle when parents finally achieve a degree of independence, from both the burdens and the delights of raising children. At this stage in the life cycle, marriage often comes under strain. As in the earliest stage of marriage, the couple may find themselves alone with each other. For many it is a period of crisis, since it involves a change in roles. This may have advantages. The husband and wife can travel more freely or go out more often without worrying about their children. They may have greater financial independence if their wages or salaries are higher than in their younger days and if they do not have to spend as much money on their children, who, by this stage, probably have jobs of their own.

Readjustments of middle age

On the other hand, middle age is coupled with physiological changes that can cause health problems or simply a feeling that life is passing by. Women who have taken primary responsibility for raising their children may feel a loss at their departure. The same may apply to men, of course, but usually not to the same degree. New relationships are formed with children-in-law, and it is often difficult for parents to adjust to the fact that their children share their love with their own spouses as well as with the parents. (Mother-in-law jokes are found throughout the world and have much basis in fact.) In addition, both men and women may feel a loss in confidence owing to the feeling that they are no longer as sexually attractive as they once were. This too may put strains on the marital bond and cause partners to stray.

Old age: children caring for parents. A precise distinction between middle age and old age is virtually impossible to draw. Certainly retirement marks a commonly recognized point of transition for many people, though they may not wish to consider this as the beginning of old age. Within the family a key factor, equally important, is the stage at which the parent–child roles are reversed, when children express their concern for the well-being of their parents by providing emotional or financial support. Often a parent, after the death of a spouse, comes to live with one of the children. Sometimes a child who has become successful provides money for the upkeep of the parents, or surviving parent, even if living elsewhere.

The transition to old age is, of course, not necessarily abrupt and does not necessarily lead to a feeling of alienation, as was once supposed. On the contrary, several sociological studies, notably in the United States, the United Kingdom, and Denmark, have shown that most old people have at least one child living nearby, often only a few minutes' traveling time away. In many Third World countries, where it is customary for parents and children to reside in the same village, if not in the same household, contact between children and elderly parents is even greater. In both traditional and modern cultures, adult children take great responsibility for the welfare of their aged parents. This is true not only in rural areas but also in urban ones, and for all social classes.

Death and bereavement. The death of a family member affects not only the individuals in the family but also the family unit collectively. The precise difficulties the family faces and the manner in which they cope with them depend on particular circumstances. Nevertheless, a few generalizations can be made with respect to the kinds of problems that frequently develop.

For the family as a whole, one obvious effect is the disruption of normal activities. This in turn may lead to the surviving family members' gaining greater resolve to face up to their loss collectively. On the other hand, it may lead instead to fragmentation similar to that which follows divorce. Indeed, divorce and death often have similar effects on children. Removal of an authority figure may lead them to take on greater responsibilities in the family, or greater independence, or it may lead instead to family conflicts and a lack of family solidarity. It may also break the remaining ties between brothers and sisters who live apart.

Losing a partner

Bereavement has a profound and more individual effect, too, on the surviving spouse. This might be enhanced if the couple were married for most of their lives and particularly if they shared their last years together in isolation from their children. Studies in the United States have shown that, in general, a husband who survives the death of his wife has greater difficulty in coping than a wife who survives the death of her husband. Not only does a surviving husband have to cope with emotional loss but he also has to take on tasks such as cleaning and cooking—which in many modern societies are still usually done by the wife. On the other hand, some sociologists have argued the reverse, that a surviving wife has greater difficulty, owing in part to the larger number of women who live to an advanced age and the resulting greater difficulty they have in finding a new husband. In other societies, such as India, widows are given distinctly inferior social status to married women and have a much harder time than widows in the West. In most parts of India widows are not permitted to wear brightly coloured clothes or to look after their appearance, and attendance at certain rituals, notably weddings, is forbidden them. Indeed, it has been suggested that the practice of suttee (immolation of the widow), formerly practiced in India, was in part encouraged by the extreme hardships that faced widows in medieval Hindu society.

MARRIAGE IN MODERN SOCIETY

Legal aspects of marriage. In nearly all cultures, marriage is distinguished from courtship or living together out of wedlock by a ceremony or series of ceremonies. These often involve bridewealth, dowry, or simply the giving of gifts, by anyone, to the newlyweds. Wedding ceremonies may be civil, religious, or a combination of the two. In Great Britain, for example, those who marry in a church must undergo a short civil ceremony (the "signing of the register") at the end of their wedding in order to validate the religious ceremony in the eyes of the state. For couples who include a divorced partner, the reverse is sometimes true: the couple may undergo a religious ceremony, in which they repeat their wedding vows, in order to receive full recognition by the Church of England of the civil ceremony they have previously undergone.

Legitimation of offspring

The main purpose of the legal validation of marriage in almost every society is to provide for the legitimacy of children. Marriage in most societies is required by custom in order to give children legal recognition as members of a family or wider kinship group and to allow them to inherit property from both parents. In some modern societies, natural children born to parents living out of wedlock are also given these rights, usually upon petition in the courts, but the norm is that the right of a child to inheritance is defined by the marriage of his parents. This is even more true when it comes to succession to titles of nobility in the male line or membership in patrilineal kin groups. Without marriage, the child belongs to the mother alone and is ineligible to inherit a title or membership in a kin group. In Western society these legal requirements date back at least to Roman times, and similar rules, distinguishing the children of wives from those of concubines, were also found in biblical times.

Marital roles. Marriage is important as the accepted institution for the expression of adult sexuality. A mutually satisfying sex life is important to both men and women, although social scientists point out that marital roles involve much more than this. Romantic love is only one of the reasons people marry. Social and economic security, and indeed social pressures, can be equally important.

The "macho" image

Relations between the sexes are to a large extent culturally as well as biologically determined. The image of the "macho" male is well-known and attributed commonly to Mediterranean and Latin-American cultures. In working-class British culture, too, tenderness in a sexual relationship has been traditionally regarded as unmanly. The public image that such men wish to project is based on sexual prowess rather than on emotional intimacy. This image may even be retained after marriage if manliness is defined by how completely a man can rule his household.

In the past, women frequently took their social status from their husbands, but by the late 20th century there was an increasing tendency for women to be regarded as equal partners in marriage. The traditional norm, where women remained at home and men went out to work, has changed rapidly. As women gain status from their own occupations outside the home, they are beginning to achieve equality with men. Women's traditional sphere of influence has been the home, however, and in cultures as diverse as the Khoikhoin (Hottentots) of southern Africa and sections of the working-class population of modern Britain, women's economic authority in the home remains paramount. Even today it is not uncommon for the British husband to depend on his wife to give him spending money, even though it may originate in his wages.

Trend toward the "symmetrical family"

In a study of the family in a low-income area of London, British sociologists Michael Young and Peter Willmott found that what had previously been regarded as the typical late 19th-century family had survived into the 1950s. This type of family was centred on the economic separation of the roles of husband and wife, sometimes with both partners working and frequently with the wife sharing domestic tasks with female relatives who lived nearby. Young married women, for example, received help from their mothers in shopping, household chores, and babysitting. In further studies made in the 1970s, however, Young and Willmott documented changes toward what they called the "symmetrical family," in which kin networks had ceased to be as extensive as in the past and husbands and wives shared domestic tasks between them. Social activities, too, had become more couple-centred, as in many cases men stayed home, perhaps to watch television, rather than to socialize with their male friends. In short, at least in London, there was a development of working-class marital roles toward a pattern similar to that found in most middle-class households. Indications are that the trend is a widespread one.

Parental roles. In all societies, past and present, parents have played a major part in caring for children. Modern parents retain the vestiges of their traditional roles, but in many parts of the world they send children to nursery school, kindergarten, and then to school, thus delegating to teachers some of their traditional responsibilities for the socialization of their children. In Western societies there is a tendency toward social equality. Wealthy parents rely less than in the past on nannies to raise their children, and lower- and middle-class parents have greater access to preschool facilities than formerly. As with marital roles, there seems to be a trend toward the reduction of differences in parental roles among social classes. The trends may indeed be related. In the non-Western world, too, modernization and economic development similar to that which took place in Europe during the Industrial Revolution are now creating a situation of greater freedom and responsibility for children. The temporary absence of fathers who take jobs as migrant labourers, for example, may place teenage children in a position of responsibility over their families. At the same time, other young people in these countries often seek employment and independence in urban areas.

Paradoxically, however, from a child's point of view, Western parents are often regarded as inhibiting independence, particularly during adolescence. In most modern societies, parents show an interest in and concern for the sexual activities of their children, something they do not do in most "primitive" societies. In modern Islāmic societies and in modern India, as in some other parts of the world, parents have the duty to ensure that their children find suitable wives or husbands, and even the children recognize this. Yet in modern Western societies the practice of parental matchmaking is regarded by children as interference in their affairs. Even before marriage, they begin to assert their independence, which, arguably, was instilled in them by their parents themselves in the socialization process.

Unequal marital roles

Feminist perspectives. One effect of the feminist movement has been an increase in sociological studies directed specifically at the roles of women in society in general and in the family and marriage in particular. The American sociologist Jessie Bernard, for instance, argued that the study of marriage must focus both on husbands' and on wives' expectations and achievements. She noted that married men are generally more successful than single men in achieving their goals in life, whereas the opposite is true for modern married and unmarried women. This view contradicts that of Young and Willmott, who saw the modern symmetrical family as a step toward greater equality and similarity of roles between husband and wife. Much of the data Bernard cited was psychological: American wives are known to be much more prone to anxiety and depression than their husbands, and this, she argued, reflects the unfairness in marital roles as conceived in American society.

Ann Oakley made a similar case for Britain. She examined the changing roles of women since the Industrial Revolution and, in particular, the emergence of a new status of woman as housewife. Oakley, and many others in feminist circles, have regarded the role of housewife as degrading to women in that it prevents them from achieving economic and social independence on the same basis as men.

Another effect of the feminist movement has been a greater awareness in society as a whole of a double standard in sexual behaviour and a growing feeling that such a standard is unfair to women. Under this code of conduct it is permissible for men to engage in premarital and, to some extent, even extramarital sexual intercourse, while women are expected to remain chaste before marriage and faithful to their husbands afterward. This double standard and its inherent assumption of sexual inequality is a cause of concern for many women.

Separation and divorce. In most societies, it is possible to end marriage through divorce. Even where divorce is not permitted, separation may be allowed, either on a de facto basis or as a legally defined status. In many Western countries, marital separation is often thought to be a relatively new problem. Certainly divorce was much more difficult, for example, in early 19th-century England (when it required an act of Parliament for each divorce granted) than it is in England today.

Religion and divorce

Religion is a major factor governing the availability of divorce. In some religions, such as Islām, marriage is regarded as a contract between two parties rather than as a union intended for life and blessed by God; the Qur'ān states that it is better to divorce than to live in an unhappy marriage, and in Islām divorce is available to the man simply by the repetition of a verbal formula three times. In many Christian countries—in particular, Roman Catholic countries—divorce is difficult to obtain. Marriage is regarded by the church as ordained by God and thus as indissoluble. In Italy, for example, divorce has been permitted by law since 1970, but the church does not recognize the legitimacy of civil divorce. Thus, couples frequently separate and live apart. In fact, five years of continued separation are required before a civil divorce can be granted, and even after this period couples do not necessarily obtain a divorce. In the Orthodox churches, divorce is permitted, but, as in Roman Catholic countries, divorces granted in civil courts are not necessarily recognized by the church. Thus, in countries such as Greece or Russia or among members of the Orthodox faith in emigrant communities, couples must seek separate divorces from the church independently of civil divorces granted by government authorities.

The frequency of divorce. Divorce rates have markedly increased in many countries since World War II and in some countries have been on the increase since the early 20th century. Attitudes toward divorce have changed dramatically in this period, with the general trend toward tolerance of the practice. Although the statistics are highly variable for overall rates, a number of correlations can be drawn between divorce and other factors.

First, divorce rates are affected by national conditions. Historical studies have shown that, in general, fewer divorces occur in times of economic depression and more in times of prosperity or war. The frequency of divorce in the United States, for example, nearly doubled during World War II.

Second, divorce rates are affected by factors related to social circumstances, including ethnic group, religion, class, and economic background. Divorce rates can be expected to be higher in groups that attach less stigma to divorce than in those that attach more. The backgrounds of partners have a more complicated effect on divorce. Studies of racially mixed marriages, for example, show that these may yield specific patterns within specific cultures. One study in the United States suggested far greater stability in marriages of black husbands and white wives than of white husbands and black wives. Such differences no doubt depend on factors derived from sex roles in American society generally, or they may be related to the kinds of people who are most likely to marry outside their group.

Third, divorce rates vary according to the family cycle itself. Many studies have pointed to the fact that the longer a couple has been married, the more likely it is to remain so. Divorce rates are highest among the young, and, if a marriage survives its first few years, there is an increased likelihood that it will continue. Another factor often cited is the presence of children as a deterrent to divorce. Empirical studies have shown, however, that this factor is much less significant than commonly believed.

Divorce and remarriage

Remarriage. With the trend toward higher divorce rates has come an increase in the number of people marrying for the second, third, or fourth time. For most people, including those who have been divorced as well as those who have not, marriage is regarded as the normal way of life. People who have been divorced at least once and have then remarried, however, have a higher probability than others to be divorced again. This tendency may reflect attitudes that result from earlier experiences, but it may also reflect the possibility that a certain proportion of the population finds it especially difficult to maintain long-term relationships. Internal psychological makeup is sometimes responsible, but factors such as occupation and lifestyle are at least as important.

EXPANDING THE BIOLOGIC FAMILY

In many traditional societies, the more children a couple has, the better off is the family. Children have been valued both in their own right and for the labour they perform. In many modern societies, too, there remains a great desire to have children. While modern birth-control methods allow many couples to limit the size of their families, modern advances in medical technology enable those who cannot produce children naturally to do so artificially. At the same time, the legal mechanisms of fostering and adoption offer traditional alternatives for expansion of the family beyond biologic children.

Artificial insemination. While many couples use birth control to limit the number of children, others have taken advantage of changing mores and new technology to increase their chances of having children. One method is artificial insemination, which involves the implanting of semen into a woman's uterus by means other than sexual intercourse. The semen may come from the woman's husband or, if the husband is sterile or is suspected of carrying a hereditary disease, from a (usually anonymous) donor. Artificial insemination by donor, although still rare, is becoming more common and socially acceptable. Nevertheless, for many people there are moral or legal complications.

Many modern Western legal systems fail to distinguish the pater (social father) from the genitor (presumed biologic father), while most traditional societies do make the distinction. Where it is not made, the law may regard the semen donor as the "father" of a child produced by artificial insemination and oblige him to bear financial responsibility for the child. Artificial insemination is even regarded by some people as a form of adultery. In order to prevent such difficulties, clinics that perform the insemination keep the identity of donors secret. Often they also mix donated semen with the semen of the prospective social father, in order to preserve the legal fiction that the pater and genitor are the same man.

Surrogate motherhood. The late 1970s saw the birth of the first "test tube babies," conceived in vitro ("in glass") under laboratory conditions. In vitro fertilization normally begins with the extraction of an ovum, or egg, and the fertilization of the ovum in a laboratory dish. The fertilized ovum is then introduced into the uterus, where it develops normally. In itself, in vitro fertilization is not particularly problematic, since it can and often does involve simply fertilizing an ovum from the woman who will carry the child. It becomes problematic, however, when the woman is a surrogate mother.

Problems of surrogacy

A surrogate mother is a woman who carries a child on behalf of another woman, who will become the child's social mother. The social mother will also be the child's genetic mother if she donates the ovum. The problem is that three roles normally borne by one woman—genetic mother, childbearing mother, and social mother—are now divided between two. In addition, the sperm may come from the husband of one of the women or from another man. Because of these complications and the emotional strain on both mothers (and potentially the child), surrogacy has been made illegal in some countries. In others, it is becoming institutionalized, and laws are being changed in order to define precisely the rights and obligations of parents and children in this situation.

Fostering. Fostering is the practice of using a parent or set of parents to care for someone else's child on a long-term basis. Often the child's own parents have died or have been declared legally unfit to look after him. Modern government social services and some private agencies place such children with families they believe will give them good homes.

Adoption. Fostering is often a first step toward adoption. Although both practices involve the assumption of parental roles by persons who are not the child's biologic parents, adoption involves legal considerations not found in fostering.

The original ancient Roman notion of *adoptio,* or "adoption," was simply one of passing legal authority over an individual from one person to another, often for the purpose of making alliances and securing the inheritance of property. In Roman times the person who was adopted was most often an adult male who continued, even after his adoption, to retain ties of love and duty toward his own living parents. In modern society these ties are normally broken in favour of ties of affection between the adoptive parents and children. The modern notion of adoption, then, combines legal aspects of the Roman notion with the affective aspects of both fostering and biologic parentage.

Adopted children in most countries today enjoy the same privileges as natural children. They are treated as fully part of the family into which they are adopted. Adoption gives couples who are unable to produce children of their own the chance to raise children, who might themselves not otherwise find a home.

FORMS OF FAMILY ORGANIZATION

The nuclear family. The nuclear, or conjugal, family is the basic unit of family organization in virtually every society. It is generally defined as a married couple and their children (including adopted and fostered children, as well as the couple's natural children). Other forms of family organization, such as compound and joint families, are in a sense built upon the nuclear family or contain units comparable to it in their structure.

The primary unit of production and consumption

In many modern societies the nuclear family is identical to the typical household unit. Members of the nuclear family share the same dwelling place, usually a single house or apartment. In agricultural societies the nuclear family is often the primary unit of production, sharing tasks and taking collective responsibility for the income that sustains them. In both agricultural and other types of communities, the nuclear family is almost always the primary unit of economic consumption. One or both parents, and sometimes children, earn money outside the home and then share at least some of the fruits of their labour with the family as a whole.

The one-parent family. A common variant of the nuclear family is the one-parent family. This form consists of one parent and his or her children. One-parent families may be formed through widowhood, divorce, or separation. They may also be formed when an unmarried person, usually a woman, raises children on her own. In many Western industrialized societies, the one-parent (especially the single-mother) family is becoming more common and tolerated. However, the extent to which it is as successful as the traditional nuclear family is a matter of conjecture.

In many traditional cultures an unmarried mother is encouraged or even forced to marry, or else she is required to give up her child for adoption by another family. It is becoming increasingly common, however, for a mother to retain her children and raise them, often with the help of her own parents or of government social agencies. In many countries this type of arrangement is more socially acceptable than formerly. In some parts of Africa, for example, female-headed one-parent families are actually more common than nuclear families as the basic household unit.

Sometimes such an arrangement is permanent, but similar female-headed households are also common in places where men are forced to leave their wives in order to find work. In southern Africa migrant labourers often have to leave their families for years at a time. This variant, however, is regarded by many sociologists as a form of nuclear, rather than one-parent, family organization. The man supports his family with wages earned away from his marital home but continues to regard it as his home, even though he may live for extended periods elsewhere. A somewhat different but related form of family organization is the West Indian matrifocal family (see below *Universality of the family*).

The compound family. The compound family is common in many traditional African societies, as well as elsewhere in the world, and is found especially where polygamy is permitted. The compound family consists of a central figure (normally the household head), his or her spouses, sometimes concubines, and their children.

Polygamy

In West Africa, for example, a man may have several wives, with each wife as the head of a subfamily unit composed of herself and her children. The wives may occupy separate dwellings within the same compound or homestead, and they often cooperate in activities such as gardening or in raising children. Children may address each woman as "mother," but they know which is their true mother and to which subfamily they belong. This form of family organization can be seen as an overlapping set of nuclear families, each with the same man as family head. While male-headed compound families are far more common, a female-headed, polyandrous version (one woman married to several men) is traditional among the Todas of South India.

The joint family. A much more common form of family organization, one that is typical in parts of India, is the joint family. In this type, a group of brothers and their wives and children all live together in the same household. In India the joint family may have property held in common by male members of the small, coresident patrilineal kin group, though joint families in other parts of the world (traditionally in China and in parts of Africa) do not necessarily have such collective rights. Sometimes the full joint family is a phase through which nuclear families pass, dividing upon the death of key persons and forming again as men are able to incorporate their sons' wives and children into their own household.

If one person is designated as the family leader and another (normally his eldest son) as his successor, this is often described as a stem family. Such arrangements occurred in feudal Japan, in feudal and postfeudal Europe, and among some groups of immigrant farmers in the United States. Joint and stem families do not have to be based on a patrilineal kin group. Sometimes, especially in Africa, they are composed of people related to one another in various ways, including blood ties through women and ties through marriage.

The extended family. There is no precise distinction between the joint family and the extended family, and the latter term can be used to encompass both. In a narrower sense, sociologists usually think of the extended family as being larger and maintaining less control over its members than does the joint family. In most extended families, the marital bonds between spouses are stronger than the kinship bonds between, for example, the brothers who are the focal people in Indian joint families.

Economic cooperation in extended families

As a household unit, the extended family is most common where ties between kinsmen are important for economic reasons. It was common in Europe during and just after the Industrial Revolution and, more recently, among European immigrant communities outside Europe. It is still common in many parts of the Third World, in both agricultural and industrial contexts, and in Asian communities in the United States, Canada, and Great Britain. Typically, a group of kinsmen live together and share resources. In a traditional context, there may be common property in the form of agricultural lands, livestock, or ritual property such as sacred objects and sacred sites. These last are important in some African societies and among the Australian Aborigines. In modern societies the extended family offers benefits in areas where government agencies have not penetrated or where essential services are not adequate or not provided. Even where there is no common property, members of extended families may draw on one another when they need financial help.

The term extended family also applies to family units that do not establish a single place of residence. The case mentioned above, of working-class people in London who at least into the 1950s shared domestic chores, is one example. The key aspect of the extended family is not that it necessarily shares the same dwelling or place of residence but that relatives give material assistance to one another and share economic responsibilities.

Kin networks. As extended families disperse and government agencies take over economic responsibilities formerly held by them, extended families become kin networks. This has happened in most modern societies. Whereas the extended family is usually associated at least with residential proximity, if not coresidence, kin networks for many people stretch around the world.

An interesting result of worldwide migrations, such as those from Europe during the late 19th and early 20th centuries, or following World War II, has been the extension of kin networks. Far from being dissolved, as some people suppose, these networks often take on a peculiar significance. Many Americans, Australians, New Zealanders, and others keep in contact with their countries of origin in Europe, Asia, and elsewhere. This contact contributes to their personal identity with their countries of origin. Kin networks are often better maintained between dispersed families than between those in closer proximity. Australians, for example, generally know their personal family histories and keep in touch with distant relatives in the British Isles and elsewhere, whereas their British cousins tend to have less interest in these matters and may have less need to keep in touch with relatives within the British Isles.

A less dramatic but equally important form of kin network is that between urban and rural areas within a particular country. As migrations to urban areas have increased, contacts have been kept up, and urban families may keep in touch with relatives in cities other than their own, as well as with relatives in their places of origin. This happens less in Europe, but it remains important in many parts of North and South America, the West Indies, Africa, and Asia.

UNIVERSALITY OF THE FAMILY

Murdock's hypothesis. In 1949 the American anthropologist George Peter Murdock published the results of a major survey of kinship and social organization in a worldwide sample of 250 societies. Murdock's starting point was the family, and on the basis of his survey he argued that the nuclear family is universal, at least as an idealized norm.

Universality of the nuclear family

All of the societies in Murdock's sample exhibited some form of family organization. More specifically, although many societies were organized into polygamous families and extended families, even these had as their basis at least two nuclear families per polygamous or extended family household. The polygamous (compound) family was made up of two or more nuclear families affiliated through plural marriage, while the extended family consisted of two or more nuclear families joined together through parent-child ties. In Murdock's sample, 47 societies had only the nuclear family level, while 53 possessed polygamous but not extended families, 92 had some form of extended (including polygamous-extended) family organization, and the remainder proved impossible to categorize on the basis of information available at the time. Murdock's key point was that, even where complex forms of family organization occur, nuclear families are still found as the basis of the more complex forms.

Murdock argued further that the nuclear family is not only universal but also universally important. Earlier writers had argued that in many tribal societies the nuclear family is insignificant and serves no important functions in the lives of most people. Murdock, in denying this view, pointed out that the key functions of the nuclear family and its universal status are most apparent when viewed in reference to the relationships that make it up. The key functions include the sexual, economic, reproductive, and educational aspects of the family. The relationships include the bonds between husband and wife, father and son, father and daughter, mother and son, mother and daughter, brother and brother, sister and sister, and brother and sister. These eight relationships have come to be known as those of primary kinship, and they are normally the relationships through which all more distant ties of kinship are traced.

The Nāyar case. The Nāyar family evokes interest because it casts doubt on the universality of both the nuclear family and the institution of marriage. The Nāyars are a high-caste group in southwestern India. Modern Nāyar families are not appreciably different from those of other Hindu groups, but before around 1792, when the British assumed control over the area where they live, Nāyar family life was very different.

According to a number of scholars who studied 18th- and 19th-century reports on their social organization, marriage did not exist among the Nāyars, although certain customs that bear a resemblance to aspects of marriage did. In particular, these included the tali-tying ceremony and legitimate unions between a woman and a series of lovers or "husbands" known as *sambandham* partners.

A tali is a gold or silver emblem that, in other parts of India, is tied around a woman's neck by her husband during the wedding ceremony. Among the Nāyars it was tied instead by a man of equal or higher status, sometimes a non-Nāyar, on a Nāyar girl during a ceremony that otherwise resembled more an initiation rite than a marriage. Several girls received talis at the same time. Some Nāyar girls removed their talis soon after the ceremony (which would never be done elsewhere in India), and in no case did the Nāyar tali-tying ceremony imply an enduring sexual relationship between the girl and her tali-tier.

In contrast, the *sambandham* relationship involved no religious ceremony, but it did involve a sexual union. Each woman took a series of partners through her life. She could, in fact, be involved in more than one such relationship at a time. (The explanation for such an arrangement may lie in the fact that the Nāyars were traditionally a warrior caste, women being left alone to look after their households and children while the men went to war.) Apart from gifts to his partners, a man had no obligations within the *sambandham* relationship. His only strong ties were to the family in which he grew up, which included his mother and other relatives related through his mother, such as his sisters and brothers. The father was not socially important, and a man had no obligations toward his children. Nevertheless, he did have obligations, through his female relatives, to a kin group including his mother, mother's mother, mother's siblings, and sisters' children. His responsibilities were to his sisters' children, not his own, and his sisters' *sambandham* partners' responsibilities were to their sisters' children.

It is doubtful that the term nuclear family accurately applies to this arrangement. Some scholars use the term subnuclear family, which retains the notion of family organization, for such an arrangement, and indeed the traditional Nāyar subnuclear family bears some resemblance to the one-parent family in Western society. The Nāyar system can also be regarded as separating the two phases of Hindu marriage and two or more of the roles normally ascribed to a Hindu husband. Among other Hindus (and indeed among the Nāyars today), the tali-tier and the lover are the same person, whereas in the past the Nāyars held these two roles to be distinct.

The West Indian matrifocal family. The West Indian matrifocal family is another well-known form of family organization that casts doubt on the universality of the nuclear family. For many lower-class West Indians, both rural and urban, the role of the father in family life is negligible. The mother is the central figure, hence the term matrifocal (meaning focused on the mother). In the usual pattern a household comes into existence after a man and a woman set up house together. Their cohabitation is sometimes based on a legal marriage, but this is not necessarily the case. When children are born of the union, they are looked after by their mother, who in turn depends on her husband or lover for financial contributions toward running the household.

Minimal role of the husband

What makes the matrifocal family unusual is that the husband takes little or no part in child care and may indeed spend little time at home, often living elsewhere in the same community. Although in other parts of the world such behaviour would be frowned on or even thought of as deviant, in the West Indies it is socially acceptable. Eventually the older children, when they leave school, contribute toward the earnings of the family, and the importance of the father may be reduced even further. From this point on the mother is not only the focus of emotional ties but also the centre of economic and decision-making activities for the family. This is true whether or not her husband or lover is present as a member of the household. Older girls frequently take lovers and have children of their own while still living with the family. These children, in turn, often grow up looking to the focal figure of the family, their maternal grandmother, as the dominant figure in their lives.

The matrifocal family often dissolves with the death of the focal figure. Sometimes a mature daughter, with the help of her father, is able to keep the family together for a time, but this is not usually the case. The brothers and sisters normally move away to set up their own households, and they repeat the cycle with matrifocal families of their own.

This form of family organization, now common in the West Indies, bears no necessary relationship to the family's economic needs, but its origins may ultimately be economic. It has been argued, for example, that the female-headed household is descended from the separation of men from their families during the period of plantation slavery. Another view places the origin of the custom even earlier, in the West African compound family and the practice of polygamy (see above *Forms of family organization*).

The Israeli kibbutz. A kibbutz (plural kibbutzim) is a type of agricultural collective found in Israel. Its typical features include the collective ownership of property, communal living, and the rearing of children by the community as a whole rather than by their parents alone. Although there are differences between kibbutzim, for example, in religious belief or in the degree of social ownership, the structure of kibbutz society in general has frequently been proposed as a counterexample to the view that the nuclear family is universal.

Murdock argued that the nuclear family in all societies performs sexual, reproductive, and economic functions. In the kibbutz it is the case that sexual and reproductive functions are served through marriage. After a period of cohabitation, kibbutz members normally marry under Israeli law, which is necessary in order to grant legal rights to their children. Yet contrary to Murdock's definition, the relationship called "marriage" in many kibbutzim has no economic functions. Economic activities such as working in the fields or with agricultural vehicles, and even activities like sewing, laundering, and cooking, are performed for the whole of the kibbutz. Women do not change their names upon marriage, and they continue to work as before. Meals are taken communally and not in a family unit.

Communal raising of children

Education, too, is often the responsibility of the kibbutz as a whole. But whereas this is true to some extent in all modern societies in which children attend school, the kibbutz takes the principle a step further. In many kibbutzim, children are raised from a young age by nurses and teachers, not by their parents. They eat and sleep in special quarters, not in the marital quarters of their parents. The purpose of these arrangements is to instill in children at a young age the communal values of kibbutz life. One interesting side effect, however, is that children brought up together in the same kibbutz tend to form sexual bonds in later life with people from outside the kibbutz. Members of their own kibbutz are all, in a sense, their "brothers" and "sisters." Similarly, although parents are much more attached to their own offspring than to those of other kibbutz members, they nevertheless refer to all of them as "our children." The structure of kibbutz life thus raises questions about the universality of the family and the psychological and sociological nature of family relations.

The traditional Nāyar family and the West Indian matrifocal family thus represent unusual systems of family organization—not because there are no cases of one-parent families or "uncaring fathers" in other societies, but because in these two systems the idea of a family in which the father plays little or no part is institutionalized as a social norm. Communal families such as the Israeli kibbutz are significant because they deny the importance of the nuclear family within societies in which nuclear families are considered normal and appropriate. Although Murdock's hypothesis may in the strict sense be overturned by these examples, they are nevertheless exceptions proving the rule that in human society the nuclear family is indeed almost universal.

Kinship

Kinship is a socially recognized relationship among people connected by marriage or common ancestry. Kinship systems are universal throughout human society, differing among cultures in their importance in the broader social structure, the number of relatives they include, and the demands they place upon the members.

The study of kinship began in the 19th century with what have been called conjectural histories—attempts by such people as the German Socialist philosopher Friedrich Engels to speculate on the origin and development of kinship systems. In the early 20th century Sigmund Freud expanded his psychoanalytic studies to speculate on the historical roots of the family, and later in the century sociobiologists used genetics and evolutionary theory to the same end.

Engels, Freud, and the sociobiologists are the best-known and among the most dramatic of those who have touched upon the question of kinship in human society. All three attempt to explain the origins and evolution of kinship and to account for aspects of kinship found universally in human societies. None of these theories, however, belongs to the mainstream of social or cultural anthropology as it is practiced today.

Most modern anthropologists deal with more specific theoretical aspects of kinship. Their interests lie mainly in explaining particular systems or particular aspects of kinship, rather than the origins, evolutionary schemes, and universal aspects of kinship. Their inquiries are both specific, to explain particular systems, and comparative, to explain the range of variation among systems. Murdock's approach to the definition of the family—although it does involve universals—is nevertheless an example of an approach based on a limited but important comparative question (see above *Universality of the family*).

Broadly speaking, current kinship studies consist of three main areas of interest: kinship terminology, descent theory, and alliance theory. Whereas some scholars treat these as distinct and competing approaches, many regard them as complementary.

ENGELS' THEORY

One year after the death of Karl Marx in 1883, Engels, his collaborator, continued the historical materialist approach to the family with a major work called *The Origin of the Family, Private Property and the State*. The book was based heavily on the work of the American lawyer and anthropologist Lewis Henry Morgan—in particular, Morgan's *Ancient Society* (1877), which emphasizes the importance of private property in the development of and changes in family structure.

Origins in primitive communism

According to Engels, the family and human kinship originated from a stage of primitive communism. In this stage, mankind was composed only of hunter-gatherers, people who made their living by foraging and had no agriculture or domesticated animals. Human society consisted of primeval promiscuous "hordes," and people mated indiscriminately with their brothers and sisters. Eventually, kinship came to be reckoned in the female line, because, with such promiscuity, a child did not know who its father was but knew only its mother's identity. Women, according to this theory, held authority over the family.

Groups of men, in Engels' theory, sometimes captured women from other hordes. As mankind advanced, mating between brothers and sisters was forbidden and male warriors were forced to take their brides from adjacent groups. In time, successful groups of males acquired many wives. Patriarchy (authority of the father) replaced matriarchy (authority of the mother) as the condition of human social life. Men might have several wives and concubines who bore children for them, and these children in turn contributed labour for the extended family group.

Monogamy, in Engels' view, came about along with an increase in private property, as men needed a family to which they could pass their inheritance. They could have sexual relations with other women, but they needed to have only one wife in order to make certain that their property would be passed to legitimate heirs. This, according to Engels, explains how the family came into being. The state then grew to enforce the laws of monogamous marriage and the distribution of private property.

Engels' theory hinges on the acceptance of the view that all mankind progressed through the same stages of evolution. In spite of its internal logic, it is accepted only by a small minority of anthropologists today. It was the officially sanctioned theory in the former Soviet Union. Many Western feminists find its explanation of early female authority attractive and convincing. But the majority of scholars believe that it is impossible to describe the origins of the family on the basis of available scientific evidence.

FREUD'S THEORY

Another historically important theory of kinship is that of Freud. He argued that parallels could be drawn between the psychological attitudes of "primitive" adults and those of European children.

The incest taboo

Freud's concern was with the origin of the incest taboo, which is practically universal in human society. He believed that children have a secret desire to commit incest and, in particular, that baby boys experience innate sexual desires toward their mothers. Noting that there is no human society in which such behaviour is sanctioned, he argued that prohibitions against incest were invented in order to combat these tendencies, which would otherwise be socially disruptive. The strictest prohibitions against incest and against marriage with various categories of kin are found, in fact, in what Freud regarded as the most primitive societies. Among Australian Aborigines, for example, a person is permitted to have sexual relations with or take as a spouse only someone from a relatively narrow category of people. Relations with anyone else are regarded as incestuous. In other so-called primitive societies, there are rules governing physical and social contact to such an extent that even behaviour thought of as quite innocent elsewhere, such as looking someone in the eye, is defined as verging on incest. These practices make sense in the context of the social organization of these peoples, but to Freud they were more than this. They illustrated the lengths to which human society must go in order to suppress the incestuous desires he considered natural.

The most speculative aspect of Freud's theory was his position on the origin of the taboos. He concluded that human society must have begun with the invention of rules to stop mating between close kin. Originally there was the primeval horde, which was ruled over by a jealous father. The father had control over and sexual access to all the females of the group, and as his sons grew up he drove them out of the horde. Then, according to Freud's theory, the sons joined together and overcame their father, killing and eating him. As a result of the guilt they felt for committing such a terrible deed, they developed a taboo on the killing of a totem animal, which represented their father. This allowed them to assuage their initial guilt, but the brothers now became rivals of one another, just as before they had been rivals of their father. In order to solve this problem, they invented a second taboo—the taboo on mating within the horde. Henceforth they were required to mate only with members of other hordes.

This idea of the origins of the incest taboo is today taken literally by hardly anyone. Yet it, and similar theories both before and since, have fascinated generations of scholars and led to much research on the question of incest in the marriage rules of different societies.

SOCIOBIOLOGY

Sociobiology originated in the 1970s, becoming established particularly after the publication in 1975 of the American biologist Edward O. Wilson's book *Sociobiology: The New Synthesis*. Sociobiology is an interdisciplinary approach combining biology and the social sciences. Its adherents argue that animal and human behaviour should be studied in conjunction with Darwinian evolutionary theory. They see anthropology as merely a subdiscipline of zoology, and in their view human kinship should be studied in the same way that zoologists study animal behaviour.

The main proponent of sociobiological theory for the study of the family has been the American sociologist Pierre L. van den Berghe. He claimed that human family systems developed as part of a complex interaction between genetic and environmental factors. The genetic factors are not only those that differ between individuals or human groups but also those that are common to mankind generally. Van den Berghe argued that human culture is not merely what is left after everything determined by biology, but rather that culture itself is an outgrowth of natural selection. The widespread occurrence of the nuclear family, the cultural rules of incest avoidance and marriage, and other aspects of kinship are, in van den Berghe's view, products of biologic evolution distinguishing mankind from the apes. Elements of fictive kinship, such as the honorary kin statuses of godparenthood and blood brotherhood, can be seen as attempts by human groups to extend kinship through culture. According to this theory, biologic kinship lies at the root of social behaviour in all human societies.

Influence of biologic evolution on culture

Some scientists have taken this biologic determinist view to yet greater extremes. The American zoologist Robert L. Trivers argued the "nepotism hypothesis." In this theory animals and humans alike are biologically conditioned to sacrifice themselves for the good of other close relatives, so that their own genes may be passed on by these close relatives when they mate. This view can explain the fact that in some societies, such as the Berbers of Morocco, it is considered appropriate and desirable to marry close kin. According to the sociobiological theory, such customs reinforce unconscious desires of self-perpetuation. The reasons given by people in such societies—maintaining close ties with relatives or keeping property in the family—do not, however, generally coincide with the unconscious desires postulated by the sociobiologists, and, for this reason, most anthropologists do not subscribe to sociobiological theory.

KINSHIP TERMINOLOGY

The study of kinship (or relationship) terminology concerns the way people in a society classify their relatives. Many scholars are interested in the social rather than the purely linguistic aspects of these classifications. How is terminology related to membership in descent groups? Which categories of relatives are permitted as marriage partners? Can generalizations be made about the correlation between terminology and social structures? Other scholars, and especially those with a training and interest in linguistics, are more concerned with the formal properties of the terminology itself. Does a given language "merge" parents with parents' same-sex siblings—in other words, call the father and father's brother by the same term? Does it "skew" generations, perhaps by calling every male member of the father's group by the term father? The scholars who address these questions often argue that terminology is independent of social structure, a school of thought that is most common among North American anthropologists.

The significance of terminology. It has long been known that languages classify the world differently. A word in one language does not necessarily have an exact equivalent in another. The way people classify the world reflects the way they think, or, conversely, they think according to the way they classify the world. That the Latin language classifies the father's brother by one term, *patruus*, and the mother's brother by a different one, *avunculus*, reflects the way ancient Roman family life was organized. For English speakers, who use only one term, uncle, for both, the distinction between these two is unimportant; an uncle on the father's side of the family is treated in much the same way as an uncle on the mother's side. The Romans, however, treated them differently, the *patruus* being a stern figure much like the father and the *avunculus* being literally an "avuncular" figure, likened somewhat to a grandfather, who unlike the father was not a figure of authority.

Similarly, the English language distinguishes some categories that other languages do not. English speakers have two terms for the other children of their parents (brother and sister) and another term for the children of their parents' brothers and sisters (cousin). Polynesian languages, on the other hand, have no equivalent of the term cousin; cousins are called by the same terms as brothers and sisters. The fact that speakers of English make this distinction reflects the fact that they treat brothers and sisters differently from cousins, or at least that they regard them as being in a different kind of relationship.

The first person to study the problem of kinship terminology in a scientific manner was Lewis Henry Morgan. Before writing *Ancient Society*, Morgan had discovered that the Iroquois Indians of New York state classify their cousins differently from English-speaking Americans. A male Iroquois calls his sisters and the daughters of his father's brothers and of his mother's sisters all by the terms *ahje* (if they are older than he is) and *kaga* (if they are younger). Yet he calls the daughters of his father's sisters and of his mother's brothers by a different term, *ahgareseh*. He makes similar distinctions between males of his generation, while female Iroquois also employ a comparable, though not identical, classification. The Iroquois traditionally behave toward all these categories of kin according to their classification. For example, an Iroquois can marry a cousin classified as an *ahgareseh*, but not a cousin classified as an *ahje* or *kaga*.

Lewis Henry Morgan and Iroquois terminology

At first Morgan thought the Iroquois were unique, but as he became more familiar with the customs and languages of the North American Indian tribes he realized that this was not the case. Eventually, he sent questionnaires on the subject to all parts of the world, mainly to American consular officials and missionaries. Morgan asked them to fill in the questionnaires with the terms for a wide variety of specific genealogical positions in all the languages the respondents encountered. The results were eventually published under the title *Systems of Consanguinity and Affinity of the Human Family* (1871). This massive work concluded with a discussion of the theory that kinship terminologies reflect preexisting social structure and that one can therefore study the prehistory of society by analyzing known kinship terminologies, an idea that became the basis of Morgan's later book, *Ancient Society*.

Although most anthropologists no longer agree with Morgan's theory, they nevertheless acknowledge the great importance of his discovery of the diversity of kinship terminology structures. Anthropologists now study kinship terminologies, or relationship terminologies (as they are variously known), in relation to existing social institutions, rather than as clues to the past. If a people classify relatives in a particular way, the implication is that they do this for a reason that may be found in their existing social structure. Even where terminologies are conservative and reflect the customs of the past, the categories are nevertheless clues to the perceptions of the people who use them.

Parents' generation terms. In the parents' generation, four forms of classification are found. These are illustrated here with examples of female relatives, though the male equivalents follow the same pattern.

The simplest type, found, for example, in Polynesian societies, classifies all female relatives (or all male relatives) in the parental generation by the same term. A person's mother, mother's sisters, and father's sisters are all called by a term that translates loosely into English as "mother." There is no equivalent to the English term aunt. This type of classification is known as generational terminology.

A more complex type, represented by the English language, distinguishes mother from aunts. This is known as lineal terminology, in reference to its distinction between lineal relatives (those from whom a person is descended, in this case the mother) and collateral relatives (those related through a sister or brother, in this case those classified as aunts or, more precisely, the person's father's and mother's sisters).

A different terminology, the most common in the world's languages, is the bifurcate merging type. This structure makes the distinction between parallel relatives (including lineal relatives and those related through a same-sex sibling link) and cross-relatives (those related through an opposite-sex sibling link). In this system a person's mother and mother's sisters are called by one term and the father's sisters by another. For instance, in Tswana, a Bantu language of southern Africa, a person's mother and mother's sisters are both called *mme* (loosely, but not exactly, translatable into English as "mother"), while the father's sisters are called *rrakgadi*.

Distinguishing parallel relatives from cross-relatives

The ancient Romans used slightly different but related terms for mother (*mater*) and mother's sister (*matertera*), but they sharply distinguished the father's sister (*amita*), who, in their patrilineal society, was closely associated with the kin group of the father. The equivalent terms for male relatives of this generation, as discussed above, were *pater, patruus,* and *avunculus,* the last being derived from *avus,* meaning "grandfather." Although the Latin terminology can be considered bifurcate merging—because relatives on the same side of the family are called by linguistically related, if not identical, terms—strictly speaking the Roman terminology is bifurcate collateral. It "bifurcates" by employing different terms for the father's and mother's sides of the family, but it is "collateral" in that it distinguished lineal relatives from collateral ones by calling the mother's sister, for example, by a different term from the mother. Scandinavian languages, as well as Old English and other Germanic languages, have had kinship terminologies of this type, with no equivalent of the modern English terms aunt and uncle. Instead, relatives are literally called "mother's sister," "father's sister," and so on.

Own generation terms. The classification of terminologies by terms for relatives in a person's own generation (brothers, sisters, and cousins) is more complex. The most prevalent classification is that of George Peter Murdock, who distinguished six types.

Murdock called the simplest type of terminology "Hawaiian." In this type, often found in societies that have a generational-terminology structure for the parental generation, there is no distinction between sisters and cousins; all are termed "sister." Similarly, all the males are called "brother," both by one another and by their "sisters." The term Hawaiian refers to a terminology structure like that found in the Hawaiian language, but it is not peculiar to the Hawaiian language or people. Hawaiian terminologies are also found in other parts of Polynesia and commonly in West Africa.

The "Eskimo" terminology of European societies

The "Eskimo" type is found in English-speaking societies as well as among Eskimo or Inuit groups. The formal definition of an Eskimo terminology is simply that it distinguishes sisters and brothers from cousins. Most European societies have terminologies of this type, as do small-scale hunting and gathering societies such as the !Kung of southern Africa and most (though not all) Eskimo groups in Canada, Greenland, and Alaska. It tends to be found in societies that have cognatic descent systems, that is, those that lack either strong patrilineal or matrilineal principles.

"Iroquois" systems, on the other hand, are generally found in patrilineal and matrilineal societies and in those societies that permit marriage to cross-cousins. This type, which is the most common throughout the world, is structurally related to the bifurcate-merging type. It distinguishes cross-cousins (father's sisters' and mother's brothers' children) from parallel cousins (father's brothers' and mother's sisters' children), and it often classifies parallel cousins by the same terms as brothers and sisters. English-speaking anthropologists have had to invent the words parallel cousin and cross-cousin in order to talk about this distinction, since English speakers do not classify their own kinfolk in this way. Iroquois systems are found commonly among North American Indians, in African societies, in some Asian societies, and in other parts of the world.

Some scholars use the term Dravidian (from the name of the South Asian language family) to describe systems similar to the Iroquois but in which terms for father's sister and mother's brother are identical to those for parents-in-law. The terminologies reflect the fact that a cross-cousin in these societies is considered a person's ideal spouse. The parents of cross-cousins are considered a type of in-law, even if one marries someone else. Terminologies with this feature are found not only in South Asia but also among South American Indians and Australian Aborigines.

Two more complex types are those to which Murdock referred as "Crow" and "Omaha." These are almost invariably found in strongly unilineal societies. As in Iroquois terminologies, cross-cousins are distinguished from parallel cousins, but, in addition, cross-cousins on one or the other side of the family are equated with their parents. Crow terminologies classify the father's sisters' daughters by the same term as the father's sisters (if the society is matrilineal, these people are members of the same matrilineal group). Omaha terminologies classify the mother's brothers' sons by the same term as the mother's brothers (who are all members of the same patrilineal group).

In societies using the Crow and Omaha terminologies, many other relatives are classified by terms that similarly transcend generational distinctions. For example, among the Trobriand Islanders, the term *tabu,* for "father's sister" and "father's sister's daughter," in fact refers to all female members of a person's father's matrilineal group. Male members of the kin group are all termed *tama.* This is sometimes translated loosely as "father," even though it refers not only to a person's actual father but also includes the father's brothers, the father's sisters' sons, and even the father's sisters' daughters' sons.

The Trobriand Islanders

Such systems, found in North America, Melanesia, and Southeast Asia, emphasize lineage membership over generation or genealogical distance. Genealogical distance is a key feature specifically of Eskimo terminologies like the English one, and generation is a key feature of most other forms of kinship terminology. The Crow and Omaha terminologies, however, show that neither genealogy nor generation is universally important for kinship organization.

DESCENT THEORY

Descent theorists are more concerned with groups than with terminology, a theoretical interest that derives from the British tradition of functionalism, which dominated anthropological thinking in Britain and most of the Commonwealth from the 1920s to the 1950s. Functionalists such as A.R. Radcliffe-Brown saw societies as being made up of component parts—institutions (like marriage, chieftainship, or the stock market) and systems (like kinship, politics, or economics). Descent theorists take a functionalist view in their appraisal of the significance of group structure. In descent theory the mechanisms of recruitment to groups, and the social functions such groups perform, are the primary foci of study.

Patrilineal descent. Systems of patrilineal descent are widely distributed. The ancient Greeks and Romans traced descent patrilineally, as do contemporary societies in many parts of Africa, Asia, and the Pacific.

The defining feature of a patrilineal descent system is that membership in a social group is determined by descent through the father. A patrilineal descent group, such as the Greek phratry or the Roman gens, thus includes a person's father, father's father, father's father's father, and so on. In addition, the child of any male member of the group, regardless of the child's own sex, is a member. Thus, a person's father's brothers and sisters (all children of the father's father) are also members of the patrilineal group. Similarly, a man's children are members of his patrilineal group, but a woman's children are not members of hers (the one she was born into); they belong to her husband's group. A woman's own status as a member of her natal group or of her husband's group depends on which such membership the society recognizes.

Matrilineal descent. Matrilineal descent systems are less common than those of patrilineal descent, but they are, nevertheless, found in many widely differing societies in Africa, Asia, and the Pacific, as well as in Amerindian societies. Examples include the Trobriand Islanders of Melanesia, the Crow and Iroquois Indians of North America, the Bemba of Central Africa, and the Nāyars of India. Nāyar society, however, is unusual in that the social role of the father is virtually nonexistent (see above *Universality of the family*). Normally, matrilineal societies maintain family relationships much like those of any other society.

Distinguishing matrilineal descent from matriarchy

Matrilineal descent is defined as descent through the mother. This does not necessarily or even usually imply that a matrilineal group is matriarchal, with authority in the hands of the mother or females, but only that a person traces membership in the group through female links. Authority within the family or kin group may be in the hands of the father or, more commonly, in the hands of the mother's brother. The mother's brother is a focus for the kin group because, for any given person, the mother's brother is the closest senior male in the group. The father is a member of a different matrilineal group.

A matrilineal descent group includes a person's mother, mother's mother, mother's mother's mother, and so on, as well as the descendants of all these people in the female line. In a matrilineal society a person's mother's brothers and sisters, and his own brothers and sisters, are all members of such a group. A woman's children are members of her group, but a man's children are not members of his; they belong to his wife's group.

Double unilineal descent. Double unilineal, or duolineal, descent is very rare. Arguably, a form of double descent exists among groups of Australian Aborigines, but the most definitive examples are found in Africa. The Yakö of Nigeria and the Herero of South West Africa/Namibia and Botswana are the best-known. The principle of double descent is that two kinds of descent group, patrilineal and matrilineal, exist simultaneously in the same society and that each person belongs to both. Often the two groups have different functions. Among the Yakö, for example, residential groupings are patrilineal and land is inherited through the father, while movable property is inherited within matrilineal groups. A person has obligations toward each of the kin groups to which he belongs.

Double descent is similar to, but distinguished from, complementary filiation. Complementary filiation occurs in patrilineal or matrilineal societies when a person has obligations toward kin on the opposite side of the family from which he traces descent. In this case, however, only one kind of descent group is recognized, either patrilineal or matrilineal, not both. The Tallensi of Ghana, for example, are patrilineal, but in this society an individual has obligations not only to his own patrilineal group but also to his mother's patrilineal group.

Cognatic descent. Cognatic, or bilateral, descent is, in a sense, the opposite of double descent. In a cognatic society there are no unilineal groups (*i.e.*, groups descended strictly in the father's or mother's line). A person is reckoned to be equally related to kinfolk on either side of the family. Western societies are mostly cognatic: although surnames, titles of nobility, and so on are inherited patrilineally, there are no longer any patrilineal descent groups as such. For example, a 20th-century Italian, unlike an ancient Roman, feels no closer to his father's brother's child than to any other cousin. They share the same surname, but they do not share membership in a descent group comparable to the Roman gens.

Cognatic descent and the modern world

Most modern industrialized nations have cognatic kinship systems, and so, too, do most hunting and gathering societies. In the latter case, persons may join either their father's or mother's band or, often, the band of the spouse or some more distant relative. These bands, consisting of perhaps 25 people among most African and Asian hunter-gatherers, or up to a few hundred in the case of native North Americans, are descent groups, even though they are not unilineal descent groups.

Finally, there are some societies that recognize an ideal of patrilineal descent but in which persons may opt for tracing descent through a female link. This arrangement, known as ambilineal descent (through either line) bears some relation to the cognatic descent system of egalitarian hunter-gatherers, but it is found instead in the hierarchical societies of Polynesia. In these societies, a person may join the group that offers the most prestige, either the father's or the mother's, but in so doing the person gives up any rights held in relation to the other group.

Fictive kinship. *Forms of fictive kinship.* Perhaps the best-known form of fictive kinship is godparenthood. This is an institution found in many Christian societies, where the ritual sponsors of a child at baptism, its godparents, act as quasi-parents, promising to look after the spiritual interests of the child. The relationship between godparent and godchild is not, strictly speaking, one of kinship. Although godparents are regarded as being like parents in certain ways, they are not seen as part of the actual kinship system. Nevertheless, certain elements of the godparent relationship come very close to kinship. For example, marriage to a godchild or to a godparent's child may be forbidden. Such rules mimic those of the incest taboo and of close kin exogamy (obligatory marriage outside the group).

In addition to the relationship between godparent and godchild, relations are established between the godparents and parents. This is particularly true in certain Roman Catholic societies, notably in western Mediterranean and Latin-American countries. There, the notion of compadrazgo (as it is called in Spanish) includes fully this cluster of relationships; parents and godparents are said to be compadres, and they are required by custom to help each other in times of hardship, to lend each other money, and to offer support, for example, at festival times.

Fostering may also be regarded as a form of fictive kinship in which foster parents provide for children and give moral as well as material support. Fostering differs from adoption in that the latter incorporates the child into the family fully and thus provides for true (social) kinship rather than merely fictive kinship. The distinction is not absolute, however, because specific ideas about what is and what is not kinship differ between cultures.

Blood brotherhood

Fictive kinship also includes blood brotherhood and other institutions in which people maintain a special, but not quite a kin, relation to one another. Among various African peoples, for example, bonds of blood brotherhood unite individuals in formal ties that are invoked in times of need. The Azande of Central Africa initiate such ties by a ritual in which each party swallows some of the other's blood. Other African peoples initiate the bond by mixing blood directly in wounds cut for the purpose. Such relationships sometimes unite not only individuals but also groups, as for instance among the Chaga of Tanzania, where blood brotherhood ties between chiefs establish alliances between their entire chiefdoms.

The distinction between real and fictive kinship is not precise but depends on many cultural factors. The only thing all fictive kinship has in common is that some aspect of the relationship is regarded as fictive, while another aspect is regarded as true kinship.

The metaphorical use of kinship terms. In all languages kinship terms form a recognized vocabulary used to designate relatives (see above *Kinship terminology*). Yet these terms are not always used only in their literal, kinship context. They may be employed in fictive kinship contexts, as in blood brotherhood or godparenthood, or yet more metaphorically in other contexts. In politics reference is sometimes made to "Big Brother," to "brothers and sisters" in the black power movement, or to "sisters" in the feminist movement. In a religious context there are "fathers" in the priesthood and "mother superiors," "sisters," and "brothers" within religious orders. Children may address their parents' friends as, for example, "Aunt Mary" or "Uncle Bill."

In anthropological jargon these usages are said to be ones of connotation rather than signification, since they identify attitudes and behaviour and not kin relationship proper. In contrast, the *!*Kung classify everyone who bears the same name as close kinsmen as if they were relatives proper. If a *!*Kung man's sister is called Kxaru (a female name), then all women named Kxaru are his "sisters." A *!*Kung man may not sit too close to his sisters or tell sexual jokes in their presence, and of course he cannot marry them. The same rules apply to his sisters' namesakes. Such customs go further than those of "sisters" in the feminist movement, for example, because they identify "true" and not merely metaphorical kinship—at least as the *!*Kung see it. The *!*Kung believe that all namesakes are descended from the same original namesake ancestor, and in effect they treat the status of namesake as a genealogical position, like father, mother, brother, sister, son, or daughter.

Residence patterns. In some societies there are patterns of residence created and maintained by explicit rules. For example, in many hunting and gathering societies a newly married couple must reside for a time in the home of the wife's parents so that the young husband can hunt and provide game meat for them. He must prove himself a good provider before he is permitted to take his wife to his own family's place of residence or to establish a residence among a different band in his tribal territory. In many other societies the rules are not so explicit or obligatory, but it may still be regarded as appropriate for a couple to live in one place rather than another.

Residence patterns and descent groups

Three commonly found patterns of postmarital residence bear a direct relation to rules of descent. These are virilocal residence (coupled with patrilineal descent), uxorilocal residence (coupled with matrilineal descent), and avunculocal residence (also coupled with matrilineal descent).

Virilocal residence literally means residence in the locality of the husband. Sometime this is called patrilocal residence (residence with the father) because the husband's place of residence is also normally that of his father, or simply because the children of the marriage will grow up in their father's natal home. The consistent practice of virilocal residence automatically creates groupings of patrilineally related kin, each residing at the same locality. This may be a large clan territory, but more commonly it is a smaller place. Virilocal residence may be permanent or temporary. If only temporary, it may have less effect on the maintenance of patrilineal kin groups. In fact, the practice of virilocal residence is common in matrilineal and cognatic societies as well as in patrilineal ones, but, in cases where descent is not congruent with residence, it does not affect descent group organization.

Uxorilocal residence is residence in the wife's place. It is also known as matrilocal residence (residence with the mother). Just as virilocal residence keeps men of a patrilineal group together and disperses the women, so uxorilocal residence keeps matrilineally related women together and disperses the men. In some societies—for example, the Bemba of Zambia—uxorilocal residence permits a daughter to work the fields she will inherit from her mother.

An alternative residence pattern, frequently found in matrilineal societies but only rarely, if ever, found elsewhere, is avunculocal residence. The term means residence with a man's mother's brother (Latin *avunculus*). Among the Trobriand Islanders (the best-known example of this residence rule) each boy leaves his parents' marital home well before the age of marriage in order to live in the village of his mother's brother. Although he has not lived there before, he is taught to regard the village as his own because it is the village of his matrilineal kin group. His brothers, his mother's sisters' sons, and other members of his matrilineal kin group also move there when they reach the appropriate age. Girls remain in their natal village until marriage, when they move to the villages of their respective husbands. These villages, of course, are those of their husbands' mothers' brothers, rather than those in which their husbands were brought up. This practice creates residential units consisting of matrilineally related males, their wives (who come from other kin groups), and their young children. The women through whom all the men are related are forever dispersed, living first in the villages of their fathers, then in the villages of their husbands, but never in their "own" villages. Such an arrangement works well in a society such as Trobriand, in which descent is through women but authority is in the hands of men.

A fourth possibility would be a form of organization in which the female members of a patrilineal group reside together and the male members of the group are dispersed. This pattern would be generated by a rule of amitilocal residence (from *amita,* Latin for "father's sister"), which would require females to live with their father's sisters and males to move in with their wives. However, this possibility, the mirror image of avunculocal residence, is unattested in the ethnographic record.

Various other possibilities are found. These patterns do not create or maintain unilineal kin-based residential groups as do the types discussed above, but they can have an effect on building cognatic kinship groups or, indeed, in dispersing unilineal groups. The kind of residence variously known as natolocal (residence in the place where one was born) or duolocal (residence in two places) requires that each married partner remain in separate childhood homes rather than form a new home together after marriage. Although rare, it does occur in some West African societies, if only as an initial stage in married life. A somewhat similar pattern involves men living in a "men's house," with only women and children living in the family home. This kind of residence, which has no Latin-derived term to describe it, is not uncommon in Papua New Guinea and elsewhere in Melanesia.

Neolocal residence and the modern world

A much more common type of postmarital residence is neolocal, which involves moving to a new place of residence, neither the wife's nor the husband's natal home. Ambilocal residence, in contrast, is residence either in the wife's home or in the husband's, often when a couple need the material support of one set of parents before setting up a home of their own. Both neolocal and ambilocal residence are found as alternatives in most modern Western societies, but neolocal residence is generally the norm. Neolocal residence is particularly common, worldwide, in those societies in which the typical household is made up of a single nuclear, or conjugal, family.

Finally, uxori-virilocal and other combinations of the basic residence patterns create greater complexity in the fitting of residence with the cycle of family life. In uxori-virilocal residence the couple live first with the wife's group and later with the husband's. The example of a hunting and gathering society in which a husband must live for a time in his wife's group and hunt on behalf of his in-laws is discussed above. If after this period of "bride service" he and his wife go to live in his own home territory, the family cycle may be described as involving uxori-virilocal residence.

It is important to remember that the labels used in this discussion describe only the typical rules or practices of postmarital residence, and especially the ideal type in any particular society. There is great variety among the world's societies, and within each society, in residential groupings and family structures. These patterns depend not only on the ideal rule of residence but also on the size of the household, whether polygamy is permitted, the size and arrangement of the community, and many other cultural, demographic, and spatial factors.

ALLIANCE THEORY

Alliance theory emphasizes the marital bond and relations between groups. It is derived from French structuralism, in particular from the work in the field of kinship by Claude Lévi-Strauss and Louis Dumont. Structuralism is more concerned with the collective thought of a people than with their social institutions.

The English term alliance, in its technical sense, carries the specific meaning of alliance through marriage, a connotation derived directly from the French word *alliance* (meaning "marriage"). Alliance theorists pay close attention to those kinship systems in which rules of marriage between groups appear to dominate a large area of social endeavour. In particular, they analyze the rules that determine which people a person may marry and which people he may not. These rules, in turn, are based on rules of incest avoidance.

The incest taboo. All societies have a concept of incest, and all societies have a prohibition, or taboo, against it. Definitions of incest vary according to definitions of who is close kin. Reactions to violations of the taboo also vary from society to society. In some cases incest is thought of with abhorrence; in other cases, with mild amusement. The vehemence of its condemnation may also depend on which specific incestuous relationship is involved, whether the parties are children or adults, and the circumstances of the violation.

No one knows how the incest taboo originated. Theories of its origin are diverse. As discussed previously (see above *Freud's theory*), Sigmund Freud believed that people have an innate desire to commit incest and that the incest taboo prevents them from carrying out such deep-seated desires. In contrast, the Finnish anthropologist Edward Westermarck, writing some two decades before Freud, argued that human beings find naturally abhorrent the idea of sex with close family members. His theory was that "familiarity breeds contempt"; in other words, the innate desire is to avoid incest, not to commit it.

Incest taboos and exogamy

Closely tied with the incest taboo is the practice of exogamy, or marriage outside the group. However, while these matters are clearly related, there are crucial differences. First, by definition, incest involves sex and exogamy involves marriage. Second, incest taboos and rules of exogamy do not always coincide. In other words, in some societies it is not considered incestuous to have sexual intercourse with persons who are forbidden as spouses. Third, incest taboos are purely proscriptive, whereas rules of exogamy may be prescriptive as well. That is, the former rules define what is not permitted, while the latter may determine whom one ought to marry. Indeed, in many societies the rules of exogamy are stated in a strongly positive way, notably in those societies in which the social structure is closely bound up with rules of marriage to particular cousins.

Elementary and complex kinship structures. According to Lévi-Strauss, the incest taboo and exogamy lie at the root of human society. The incest taboo is on the one hand natural and universal, since every society recognizes it, and on the other hand cultural, since exactly which relatives are forbidden to marry vary widely among societies. Generally speaking, the specification of the taboo and the consequent marriage rules take two possible forms; Lévi-Strauss called these "elementary" and "complex."

Elementary kinship structures are those in which there exists a positive rule for marriage to someone of a particular kinship category, for example, to a cross-cousin (father's sisters' and mother's brothers' children) or someone of a wider category including cross-cousins. In principle, elementary structures offer limited choice of a spouse. Complex kinship structures (which, ironically, are much simpler to understand) are those that have negative marriage rules—*i.e.*, those specifying which persons one may not marry. Since ancient times all Western societies have had complex structures, because under their rules of kinship, brothers, sisters, children, and other close relatives may not marry, although a person may marry anyone else.

Modern societies in most parts of the world have complex structures—those in which the patterns of marriage are not precise or easily discernible and, hence, are "complex." Many scholars believe that these complex systems emerged from elementary ones. Certain systems fall between the elementary–complex distinction. The traditional kinship systems of some native North American and West African peoples (the so-called Crow-Omaha systems), for example, have a complex set of negative marriage rules, but they have so many such rules that the choice of a spouse is as restricted as in an elementary system. In such societies entire clans, and even clusters of clans presumed to be related, are forbidden as possible spouses.

Prescriptive and nonprescriptive systems

Since the formulation of Lévi-Strauss's theory in the 1940s, anthropologists have tried to define more precisely the essential properties of elementary structures. For the British anthropologist Rodney Needham the crucial distinction is not between elementary and complex but between prescriptive and nonprescriptive (formerly called preferential) systems. Prescriptive systems include those in which the kinship terminology defines exactly all the marriage possibilities. In some such societies the term for wife and cross-cousin is the same, whether a man actually marries one of his cross-cousins or not. The implication is simply that a man must marry someone of the category that includes cross-cousins. For a person born into a society with a prescriptive terminology, marriage to a cross-cousin is a logical consequence of the terminology structure itself.

The marriage of cousins. Elementary structures are of two types. The first type involves direct or restricted exchange, which allows the exchange of sisters as wives between groups. This system is a consequence of a rule requiring marriage to cross-cousins and more distant relatives on either side of the family who are addressed with the same term as cross-cousins. The second type involves generalized exchange, which permits a man to marry only a woman related through his mother. The closest relative he can marry is his mother's brother's daughter. A woman marries her father's sister's son or some more distant relative called by the same term as the father's sister's son. A third form of elementary structure, delayed direct exchange (involving a repeated pattern of father's sister's daughter/mother's brother's son marriages), was postulated by Lévi-Strauss, but it does not exist except as a theoretical possibility.

Systems of direct exchange are often described as "symmetrical," as cross-cousins on both sides of the family are called by the same term, and in principle a person may marry anyone of the cross-cousin category, irrespective of the side of the family or genealogical distance. Direct exchange in its most literal sense is the exchange of women as wives between groups. Many peoples with perfectly symmetrical kinship terminologies do not actually practice this pattern of exchange, however. Symmetrical systems occur in the Indian subcontinent, among most native South American peoples, and among the Australian Aborigines. The Aborigines, in particular, have extremely elaborate cosmological structures with which they classify virtually every aspect of the known universe; these structures match their kinship terminologies and rules of marriage with mathematical precision.

Systems of generalized exchange are said to be "asymmetrical," as the kinship terms for cousins on each side are different, and a man may marry only on his mother's side and a woman only on her father's. In theory, such systems widen the circle of social relationships. The men of Group A marry the women (their mothers' brothers' daughters) of Group B; but the women of Group A may not marry the men of Group B, since that would imply direct exchange or a symmetrical relationship. They must marry the men of Group C or D or E, and so on. Because of the asymmetrical nature of the relationships between such groups, generalized exchange tends to create, or at least sustain, hierarchical relations. In some societies (particularly those practicing Hinduism), it is believed that the husband's group is superior to the wife's. Thus, virtually by definition, a man's in-laws are his inferiors and a woman's are her superiors. In other cases, including many tribal societies in contemporary Burma and Indonesia, the reverse is true: a woman's group is regarded as superior, and a man's in-laws are of higher status than he is. Among certain Indonesian peoples, however, intermarrying groups are linked in a complete circle of marital alliances. Closing the circle prevents any one group from obtaining outright superiority over the others, since each group owes deference to the group from which its wives come.

BIBLIOGRAPHY. MICHAEL ANDERSON, *Approaches to the History of the Western Family, 1500–1914* (1980), discusses the main approaches and includes a useful bibliography. Important historical collections of essays are PETER LASLETT (ed.), *Household and Family in Past Time* (1972), based essentially on the demographic approach; and JACK GOODY, JOAN THIRSK, and E.P. THOMPSON (eds.), *Family and Inheritance: Rural Society in Western Europe, 1200–1800* (1976), illustrating the household economics approach. Current research in the field is published in the *Journal of Family History* (quarterly). For discussions of family and kinship in ancient Greece, see S.C. HUMPHRIES, *Anthropology and the Greeks* (1978, reprinted 1983); and W.K. LACEY, *The Family in Classical Greece* (1968, reprinted 1984). On ancient Rome, see BERIL RAWSON (ed.), *The Family in Ancient Rome: New Perspectives* (1986); and PHILIPPE ARIÈS and GEORGE DUBY (eds.), *A History of Private Life: From Pagan Rome to Byzantium* (1987; originally published in French, 1985), the first volume of a projected series that will provide coverage up to the second half of the 20th century. For the medieval period and after, see JACK GOODY, *The Development of the Family and Marriage in Europe* (1983); CHRISTIANE KLAPISCH-ZUBER, *Women, Family, and Ritual in Renaissance Italy*, trans. from French (1985); and DAVID HERLIHY and CHRISTIANE KLAPISCH-ZUBER, *Tuscans and Their Families: A Study of the Florentine Catasto of 1427* (1985; originally published in French, 1978). The most detailed study of the family from 1500 is LAWRENCE STONE, *The Family, Sex and Marriage in England, 1500–1800* (1977). Another general study, with emphasis on France, is JEAN-LOUIS FLANDRIN, *Families in Former Times: Kinship, Household, and Sexuality* (1979; originally published in French, 1976). MICHAEL ANDERSON, *Family Structure in Nineteenth Century Lancashire* (1971), is a major case study. SYBIL WOLFRAM, *In-Laws and Outlaws: Kinship and Marriage in England* (1987), examines more recent family history, especially from a legal point of view. For a philosophical treatment of the subject, see EMMANUEL TODD, *The Explanation of Ideology: Family Structures and Social Systems* (1985; originally published in French, 1983).

Essay collections on the modern family include NORMAN W. BELL and EZRA F. VOGEL (eds.), *A Modern Introduction to the Family*, rev. ed. (1968); MICHAEL ANDERSON (ed.), *Sociology of the Family*, 2nd ed. (1980); ROSE LAUB COSER, *The Family, Its Structure & Functions*, 2nd ed. (1974); C.C. HARRIS *et al.* (eds.), *The Sociology of the Family* (1979); and ROBERT M. NETTING,

RICHARD R. WILK, and ERIC J. ARNOULD (eds.), *Households: Comparative and Historical Studies of the Domestic Group* (1984). The leading journals in the field are the *Journal of Marriage and the Family* (quarterly); and *Family Relations* (quarterly), the latter devoted almost exclusively to applied studies. A detailed introduction to the sociological study of the family is F. IVAN NYE and FELIX M. BERARDO, *The Family: Its Structures and Interaction* (1973), dealing extensively with the family cycle and marriage. See also F. IVAN NYE (ed.), *Family Relationships: Rewards and Costs* (1982); and STEVEN L. NOCK, *Sociology of the Family* (1987). C.C. HARRIS, *The Family and Industrial Society* (1983), analyzes the modern family within the larger society; and D.H.J. MORGAN, *Social Theory and the Family* (1975), gives a more theoretical account. WADE C. MACKEY, *Fathering Behaviors: The Dynamics of the Man-Child Bond* (1985), is a cross-cultural study. BERT N. ADAMS, *The American Family: A Sociological Interpretation* (1971), now slightly dated, describes the subject from a sociological point of view; while DAVID M. SCHNEIDER, *American Kinship: A Cultural Account*, 2nd ed. (1980), gives an anthropological perspective. ANTHONY CLARE, *Lovelaw: Love, Sex & Marriage Around the World* (1986), based on a British television series, presents a comparative view.

A useful collection of articles on socialization in a variety of cultures is JOHN MIDDLETON (ed.), *From Child to Adult: Studies in the Anthropology of Education* (1970, reprinted 1977). JOSEPH M. HAWES and N. RAY HINER (eds.), *American Childhood: A Research Guide and Historical Handbook* (1985), is a detailed history. Modern problems are discussed in SHEILA B. KAMERMAN and CHERYL D. HAYES (eds.), *Families That Work: Children in a Changing World* (1982). A negative view of the modern family is presented in DAVID COOPER, *The Death of the Family* (1971); and R.D. LAING, *The Politics of the Family and Other Essays* (1971). PHILIP ABBOTT, *The Family on Trial: Special Relationships in Modern Political Thought* (1981), surveys the opinions of philosophers and social theorists. MICHAEL YOUNG and PETER WILLMOTT, *Family and Kinship in East London* (1957, reprinted 1986), and *The Symmetrical Family: A Study of Work and Leisure in the London Region* (1973, reprinted 1984), are important studies. Another classic study is CONRAD M. ARENSBERG and SOLON T. KIMBALL, *Family and Community in Ireland*, 2nd ed. (1968). Feminist perspectives are presented in ANN OAKLEY, *Housewife* (1974; U.S. title, *Woman's Work: The Housewife, Past and Present*, 1975); and JESSIE BERNARD, *The Future of Marriage*, 2nd ed. (1982). ROBERT S. WEISS, *Marital Separation* (1975), is based on a study made in the United States. PAUL BOHANNAN (ed.), *Divorce and After* (1970), is a collection of studies of divorce in several countries. ESTHER WALD, *The Remarried Family: Challenge and Promise* (1981), describes the topic from a family therapist's point of view. Adoption, divorce, and other aspects of family life are discussed in SHEILA B. KAMERMAN and ALFRED J. KAHN (eds.), *Family Policy: Government and Families in Fourteen Countries* (1978). Forms of family organization are studied in WILLIAM J. GOODE, *The Family*, 2nd ed. (1982). See also KINGSLEY DAVIS (ed.), *Contemporary Marriage: Comparative Perspectives on a Changing Institution* (1985). An overview of African family studies is presented in DIANE KAYONGO-MALE and PHILISTA ONYANGO, *The Sociology of the African Family* (1984). The effects of labour migration of Southern African families are discussed in COLIN MURRAY, *Families Divided* (1981). A very different geographical area is studied in RUBIE S. WATSON, *Inequality Among Brothers: Class and Kinship in South China* (1985).

The theory of the universality of the family is set forth in GEORGE PETER MURDOCK, *Social Structure* (1949, reprinted 1965). C.J. FULLER, *The Nayars Today* (1976), deals with the Nayar debate. A good discussion of the matrifocal family is R.T. SMITH, "The Matrifocal Family" in JACK GOODY (ed.), *The Character of Kinship* (1973), and other essays in this collection are also valuable. Specific problems of an Israeli kibbutz are discussed in MELFORD E. SPIRO, *Children of the Kibbutz*, rev. ed. (1975); and YONINA TALMON, *Family and Community in the Kibbutz* (1972).

Essay collections on kinship, with both comparative and descriptive studies, include PAUL BOHANNAN and JOHN MIDDLETON (eds.), *Kinship and Social Organization* (1968); JACK GOODY (ed.), *Kinship* (1971); NELSON GRABURN (ed.), *Readings in Kinship and Social Structure* (1971); DAVID M. SCHNEIDER and KATHLEEN GOUGH (eds.), *Matrilineal Kinship* (1961, reprinted 1974); and A.R. RADCLIFFE-BROWN and DARYLL FORDE (eds.), *African Systems of Kinship and Marriage* (1950). RODNEY NEEDHAM (ed.), *Rethinking Kinship and Marriage* (1971), contains a number of important papers, particularly from the perspective of alliance theory. Among introductory texts are ROBIN FOX, *Kinship and Marriage* (1967, reprinted 1983), especially good on alliance theory, though feminist critics have raised objections to its treatment of society as male-dominated; ROGER M. KEESING, *Kin Groups and Social Structure* (1975), valuable for analysis of descent and residential arrangements, though many scholars disagree with specific aspects of its portrayal of alliance structures and kinship terminologies; and LOUIS DUMONT, *Introduction à deux théories d'anthropologie sociale* (1971), an excellent treatment of both descent and alliance and of the main differences between British and French approaches to the study of kinship. Theoretical and methodological issues in the study of kinship are more fully discussed in ALAN BARNARD and ANTHONY GOOD, *Research Practices in the Study of Kinship* (1984). Another important theoretical account of the subject is DAVID M. SCHNEIDER, *A Critique of the Study of Kinship* (1984).

Most anthropological studies of kinship in particular societies are published in specialist journals or as chapters in anthropological monographs. Among book-length studies dealing mainly with kinship are JONATHAN P. PARRY, *Caste and Kinship in Kangra* (1979), on North India; BRONISLAW MALINOWSKI, *The Sexual Life of Savages in North-Western Melanesia*, 2 vol. (1929, reprinted in 1 vol., 1987), a classic study of the Trobriand Islanders; HILDRED GEERTZ and CLIFFORD GEERTZ, *Kinship in Bali* (1975); E.E. EVANS-PRITCHARD, *Kinship and Marriage Among the Nuer* (1951, reprinted 1973), on the Nuer of the southern Sudan in Africa; I. SCHAPERA, *Married Life in an African Tribe* (1940, reissued 1966), on the Tswana of Botswana; ADAM KUPER, *Wives for Cattle: Bridewealth and Marriage in Southern Africa* (1982); FRED EGGAN, *Social Organization of the Western Pueblos* (1950, reprinted 1973), on the Indians of the American Southwest; RAYMOND FIRTH, JANE HUBERT, and ANTHONY FORGE, *Families and Their Relatives: Kinship in a Middle-Class Sector of London* (1969); and KENNETH MADDOCK, *The Australian Aborigines: Portrait of Their Society*, 2nd ed. (1982). Essays illustrating the flexibility of the concept of kinship over time and place are collected in LINDA S. CORDELL and STEPHEN BECKERMAN (eds.), *The Versatility of Kinship* (1980).

FRIEDRICH ENGELS, *The Origin of the Family, Private Property, and the State* (1902, reissued 1985; originally published in German, 1884); and SIGMUND FREUD, *Totem and Taboo* (1918, reissued 1983; originally published in German, 1913), contain their respective theories. The sociobiological view, with important statements on the relation between animal and human family behaviour, is presented in PIERRE L. VAN DEN BERGHE, *Human Family Systems: An Evolutionary View* (1979, reprinted 1983); and ROBIN FOX (ed.), *Biosocial Anthropology* (1975). EDWARD WESTERMARCK, *The History of Human Marriage*, 5th ed., 3 vol. (1921, reprinted 1971), includes his theory of incest. A later discussion of the subject is found in JAMES B. TWITCHELL, *Forbidden Partners: The Incest Taboo in Modern Culture* (1987). On alliance theory and elementary structures, see CLAUDE LÉVI-STRAUSS, *The Elementary Structures of Kinship*, rev. ed. (1969, originally published in French, 1949), and his "Future of Kinship Studies," *Proceedings of the Royal Anthropological Institute of Great Britain and Ireland* (1965), pp. 13–22; as well as RODNEY NEEDHAM, "Prescription" and "Alliance," *Oceania*, respectively, in 43:166–181 (March 1973) and 56:165–180 (March 1986). For an analysis of the South Indian symmetrical system from this perspective, see ANTHONY GOOD, "Prescription, Preference and Practice: Marriage Patterns Among the Kondaiyankottai Maravar of South India," *Man*, 16(1):108–129 (March 1981).

(A.J.B.)

Family Law

In the past, family law has been closely connected with the law of property and succession, and from the records available, it must have had its principal origins in the economic and property questions created by the transfer of a woman from her father's family to the power and guardianship of her husband. Even in regard to parent and child, such legal concepts as guardianship, custody, and legitimacy were associated with family power structures and family economic interests. Family law also has to do with matters of personal status—for example, the question whether X is to be considered married or single or whether Y is to be classed as legitimate—although the incidents and importance of these distinctions often lead back to the law of property.

Family law shares an interest in certain social issues with other areas of law (*e.g.,* criminal law). One of the issues that has received a substantially increased amount of attention, from various points of view, is the very difficult problem of violence within the family. This may take the form of physical violence by one adult member on another (in this case the woman is almost always the victim), or by an adult on a child, or of some other form of violent or abusive conduct within a family circle. Difficulties can arise when the wrongdoer returns to cohabitation with the person who has made a complaint. In serious cases the only real solution may be a termination of cohabitation, or the removal of an abused child from the family unit, for example, into some form of public or foster custody. The problem is one of social importance, and some studies (*e.g.,* several done in North America) indicate that a high proportion of violent crime originates in family units.

This article is not a treatise on the family laws of the world (which would require at least a volume) but a consideration of the role of law in regard to the family and an effort to identify the main problems on a comparative basis. In recent decades, family law has been subject to reexamination in many parts of the world, and the greater legal status and independence acquired by married women has been a catalyst.

This article is divided into the following sections:

FAMILY GROUPS

A family group has a certain internal structure as well as relationships between itself and third parties. Family groups in some societies have tended to be complex, as, for example, the Roman paterfamilial group, the Chinese upper-class family, the Indian joint family, the samurai family in Japan, and many customary family structures in Africa. The family may be a part of a larger group such as the tribe or clan.

The two-parent family. At present the dominant form of the family group consists of two spouses and the children they have produced or adopted. The law, therefore, is concerned mainly with the rights and duties of husband and wife and parent and child. In a strictly monogamous society the law will forbid a husband to be married to more than one woman at the same time, while in other societies it will regulate the number of wives (as does Islāmic law). Some two-parent families are, in effect, matriarchal because the adult male is so frequently absent, as in fishing communities or among peoples with high concubinage rates. In a number of countries (although in different degrees) there is a growing acceptance of unmarried cohabitation, and there is increased concern about the legal implications of such cohabitation. Legal status

Traditionally, family law has not concerned itself much with unions that are not commenced by legal marriage, though some systems of law permit the recognition of a "natural" child by a father for such purposes as inheritance or support. The family group based on concubinage has been largely neglected by the law because such unions are often transitory and difficult to define, are considered immoral, occur mainly among poorer or less educated classes (as in some parts of Latin America), or are associated with an inferior status of the female (particularly in Asian and African countries).

The one-parent family. Where divorce and separation are common and where welfare programs are available to assist dependents, the one-parent family acquires an importance that is not adequately reflected in traditional law. It may be necessary to adapt the law to a greater extent to the needs of one-parent families in such areas as the organization of family and child welfare services and the legal and administrative machinery for family support, employment assistance, day nurseries, and the like. The head of a single-parent household may have difficulty affording the high cost of child care while working or training, especially on a modest or low income.

Even where the family is a two-parent group, it may encounter some of the problems of the one-parent family if the woman is employed outside the home. Some European countries, and the People's Republic of China, have probably gone further than most in proportions of women in employment, but many other countries also have changing patterns of female employment. In the poorer classes generally, women have traditionally worked outside the home at unskilled jobs; a newer development is the growing number of middle-class married women who are employed full-time or part-time for personal satisfaction, increased standard of living, or both.

Legal consequences of marriage. Two persons might produce the economic incidents of marriage by executing appropriate contracts or settlements. On this argument, a marriage provides a technically simple way of achieving things the parties could do for themselves in finance and property matters with greater expense and complexity. In some legal systems, in which agreement on property arrangements is central to the status of a legal wife, a contract in conventional form is the core of the constitution of marriage. The contract may be complex, with a variety of clauses, as in Muslim law or (more so formerly) as in major family unions under French or English law. In most countries today, however, the legal documentation of a marriage is mainly a registration of the event. So basically, in the legal sense, a marriage is the implied creation of certain rights or obligations such as maintenance, marital property and succession rights, and the custody of minor children.

In modern systems, the parties to a marriage can usually create the economic incidents of the marriage by a separate agreement. In some early legal systems and in present systems in which customary family law pertains, there is little choice as to the economic incidents of marriage because these are fixed by custom. In legal systems that allow substantial scope for personal independence, the spouses can take up a position of their own as to the economic basis of their family group by means of a marriage contract or a will.

Status and contract

According to the English legal scholar Sir Henry Maine, legal history shows a movement from "status" to "contract"—meaning that modern private law is less concerned with a person's class (*e.g.*, slave, illegitimate son, serf, freeman, Roman citizen) than it is with voluntary agreements among persons. A maintenance obligation has a different practical significance in a country such as Sweden, where most married women are employed and there are extensive welfare programs, than, for example, in India. Rights of intestate succession have a different social impact in a country where most people make wills than in one where most die intestate. In many countries there has been a tendency in the law to allow increasing freedom in setting up the family group and establishing its economic base. This is not true of some countries, and generally Communist legal systems impose at least a matrix of mandatory general principles designed to promote the political and social philosophy of the state. The extent of mandatory regulation, as compared with leaving decisions to the spouses voluntarily, is one of the important issues facing family law at the present time.

One feature that distinguishes marriage from a simple contract is that, in many countries, the parties cannot release themselves by mutual agreement. But some legislation in North America and western Europe comes close to permitting this; the grounds of divorce have been so widened that the marriage can be terminated, for example, after a period of separation.

The social environment. The family group is in some respects a microcosm of society, reflecting within it many facets of the whole social and legal environment. The great religions of the world have taken a special interest in family law, seeing in it opportunities for extending their moral and spiritual influence. Family law has at times been an element in promoting the interests of a particular class; in many countries there have been social hierarchies sustained by marriage and property laws. Practicing lawyers and the courts devoted much of their time to the interests of the dominant class. The family structures of the poor and the peasantry received much less attention in the law; the family law of these classes had more to do with workhouses, charitable institutions, unemployment relief, and welfare services. The 20th-century revolutions in Russia and China had as a main target the family structures of existing quasi-feudal regimes and consequently made large-scale changes in family law. The introduction of the Meiji code in Japan in 1898 similarly involved basic changes in the traditional family law, marking the emergence of a new society.

The wide differences in social organization and custom indicate that there can be no universal "best" pattern for family law. The legal regulation of familial relationships should be approached with caution lest the prejudices of a majority be imposed unnecessarily on the private lives of individuals. As the English philosopher John Stuart Mill wrote in his "Essay on Liberty," "There is a limit to the legitimate interference of collective opinion with individual independence; and to find that limit, and maintain it against encroachment, is as indispensable to a good condition of human affairs, as protection against political despotism."

The range of problems falling within the general scope of family law may conceivably increase in the future, and there are legal scholars who predict that, within the next 50 or 100 years, family law may experience some of the most critical events in the history of human societies. First, rising population in some parts of the world may bring profound changes in laws relating to the family. Methods of limiting population and of controlling reproduction and heredity will, to the extent they are put into practice, have an important effect on family relationships and the principles of family law. Second, an increasing life expectancy is itself an important influence on marriage and family structure, because it lengthens the average cohabitation until dissolution of a marriage by death. Third, new dimensions in relations between the sexes could be introduced by a developed biotechnology in the control and management of genes, cells, etc., undoubtedly raising new legal problems. Fourth, the accelerating tendency for people to group in huge urban complexes is a social force bearing on the traditional structures of family law.

CHILDREN

Laws relating to the care of children

It is almost universally the rule that natural or adopting parents have a primary duty to maintain their minor children. In the great majority of cases, the care and upbringing of a child belongs to its biological parents automatically, without regard to their qualification or suitability. No doubt this arrangement was due originally to its convenience and to lack of alternatives, although examples may be found of groups rearing their children in common (usually in tribal societies). The parental system has been justified on religious grounds. Thus, in an Irish case the court declared: "The authority of a father to guide and govern the education of his child is a very sacred thing, bestowed by the Almighty, and to be sustained to the uttermost by human law." At least one criticism of this system is the inequality of opportunity and upbringing it can offer one child as compared with another.

Common-law view of illegitimacy

Legitimacy. By the common law of England, an illegitimate child was a *filius nullius* (without relatives). There may have been two main reasons for this former, discriminatory attitude. First, certain unions between the sexes were designated as lawful marriages, and a man of importance, agreeing to his daughter's marriage, would insist on her having the status of legal wife. Second, paternity, in the legal sense, was easier to establish in the case of a lawful marriage than in the absence of a marriage. The common law of England, for example, presumes in favour of legitimacy when the child is born in lawful wedlock, even if the biological facts may be otherwise. Civil-law systems—those derived from Roman law—have been less absolute than the common law; they provide ways of legitimating a child, such as through subsequent marriage of the parents or through an act of recognition by the father. Modern statute law has brought the positions in different systems closer together and removed some of the worst features of the doctrine of legitimacy. Legitimacy is a concept of diminishing importance in modern law, and even countries that still retain it have usually modified it. They have done so by basing support obligations on parentage rather than on a legally valid marriage and by giving rights of intestate succession to children born out of wedlock. By the legal devices of legitimation and adoption and by other means, the difference between the legal status of a legitimate and that of an illegitimate child has been narrowed.

Illegitimacy is frequently associated with poverty; a contributing factor in this has been the tendency to exclude the child by law and social circumstances from what the community regarded as a customary and respectable family structure. The illegitimate child has usually lacked the financial support available in the conventional family group. Sometimes the laws have been designed to prevent the cost of support from falling on those not biologically responsible for the birth. Thus, English legislation of 1576 provided that an order could be made on the putative father for the maintenance of a bastard by the parish (the local government). The thrust of modern law-reform proposals, in general, is that the welfare of all children should be a matter of honest and effective public concern, made mandatory by statutory provisions.

Adoption. In modern societies, adoption tends to be associated with illegitimacy because adoption of the children of unmarried mothers has become common, although in some jurisdictions there is also a considerable number of married parents who offer their children for adoption. The ordinary legal principle is that the consent of a natural parent (or guardian) is required for an adoption order by a court. This consent may be dispensed with if the correct person cannot be found or has proved to be an uninterested or cruel parent. An emerging problem for family law in some countries is whether courts or public agencies should be able to effect the transfer of a child from the custody of its parents on the ground of what is best for the welfare of the child.

Adoption in Roman law

Adoption in the older legal systems (as in Roman law) was treated mainly in terms of succession law. It provided a way of introducing an outsider into a family group and so bringing him or her within the scope of the succession rules. In modern systems, succession rights and other obligations and rights in cases of adoption are usually treated by analogy with those of legitimate children; in some systems there is an explicit equation with legitimate children.

Education. The rapid development of education in the 19th and 20th centuries has had dramatic effects upon the family and upon the rights and obligations of family members. Until the latter part of the 19th century, even in highly developed countries, the organized education of children of the poorer classes tended to be casual or nil. Since then, the powers of a father to determine the upbringing of his child have declined before the advance of public education and the complex legislation and financing on which it rests. Today the pattern in some parts of the world is compulsory education up to the late teens, with extensive opportunities for higher education into the early 20s and perhaps later. Present tendencies seem to be toward even longer and further involvement with the educational process. If this trend continues, the educational systems of some countries might even become child-rearing systems providing foster homes, boarding houses, children's villages, and various health and welfare facilities. The effect on family law may be to reduce the importance of the biological link between parent and child.

Decision making. The older law in many countries regarded decision making in regard to children as internal to the family, an area in which the courts should not intervene except in cases of serious child abuse or the like. In the English common law, for example, there are decisions of the latter part of the 19th century in which this doctrine of the "family veil" was carried to considerable lengths by giving an autocratic position to the father during his lifetime, and even longer, if a testamentary guardian was appointed upon his death. In most primitive societies, customary law gave similar authority to the father, although sometimes the custody and training of girls was the special province of the mother. In modern law, the power of the father yielded to the principle that the welfare of the child is paramount; but this relaxation has raised important and difficult questions. The prevailing view is that the courts should take jurisdiction and intervene in family decision making when injustice, oppression, or cruelty might result if they did not. The consensus seems to be that it would be an extreme and undesirable principle to make parent–child relations wholly private and exclude the jurisdiction of the courts, but that it would also be extreme and undesirable to have no private domain of decision making and to bring all family disputes to court. The practical rule lies between the extremes; the application of such a rule is uncertain, and there are bound to be differences of opinion.

Traditional authority of the father

A variant of this problem is that of determining the boundaries between private decision making in a family and decision making by public authorities and services, such as child welfare officials. Here the question is to what extent the public authorities ought to share in the bringing up of children. This is a policy issue of great difficulty; in modern times, the courts have tended to intervene more than formerly in cases in which children are neglected, abused, delinquent, and so on.

Questions of custody. Cases can be cited to show the difficulty of reaching agreement among judges and others as to what constitutes the "best interests" and "welfare" of a child. The Supreme Court of Canada has supported the principle that a natural parent, such as an unmarried mother, should not be denied custody of her child merely because applicants wishing to adopt the child offer better material prospects or a two-parent instead of a one-parent household. Other views maintain that a young unmarried mother is, in general, not a suitable custodian of her newborn child and that therefore public intervention leading to adoption is usually in the child's best interests. Another disputed question is whether a difference of religion between a natural parent and an adoptive parent, or between two natural parents, is against the interests of the child. Various shades of opinion have been expressed: for example, that failure to raise a child in a particular religion may stunt its spiritual development; or that a change of religious environment could be upsetting to a child already accustomed for some years to a certain religion.

Questions of custody cannot be determined solely by deduction from a rule of law. They require the exercise of judicial discretion that takes account of all the relevant circumstances, which may be very complex. In divorce cases the situation is often a de facto one: separation of the parents has taken place some time before the legal proceedings, and the child is already in the custody of one of them, so that the divorce decree may do no more than regularize in law what has already happened to the child in fact. Some common-law courts have on occasion ordered joint custody, whereby the noncustodial spouse is involved, together with the custodial spouse, in decision making regarding the welfare and upbringing of the child. Another development of growing importance is the use of some form of family counselling in questions of custody of children. The basic argument in favour of this is that a custody plan worked out with the help of mediation and agreed to voluntarily by the parents is likely to have greater success than a custody judgment imposed on the parents after litigation.

MARRIAGE

The history of marriage is bound up with the legal and economic dependence of women upon men and the legal incapacities of women in owning and dealing with property. In Babylonian law, for example, one characteristic of a "legal wife" was that she brought property to the marriage (as a contribution to the support of the new family). In *The Story of Civilization* the author Will Durant asks what "changed virginity from a fault into a virtue" and answers the question: "Doubtless it was the institution of property. Premarital chastity came as an extension, to the daughters, of the proprietary feeling with which the patriarchal male looked upon his wife. The valuation of virginity rose when, under marriage by purchase, the virgin bride was found to bring a higher price than her weak sister; the virgin gave promise, by her past, of that marital fidelity which now seemed so precious to men beset by worry lest they should leave their property to surreptitious children."

Marriage and its relation to property

Before there was a developed law of wills, there would be a distribution of property upon the death of a paterfamilias. The property taken by a daughter on marriage may have been to compensate for her not being in the family group at the next such distribution. It was also customary for the man to make gifts to his bride, either in property or by giving services to the bride's family, as a token of the seriousness of his intentions.

Marriage as a transfer of dependence. In systems in which the females are legally and economically dependent within a family hierarchy, the juridical essence of marriage is the transfer of the woman from control by her own family to control by her husband. Marriage customs of many times, countries, and religions exhibit this principle in a variety of forms, for example, in certain kinds of Roman marriage, in marriages among the Japanese samurai, in the traditional Chinese marriage, in the Hindu marriage based on the joint family, in rabbinical law, in Muslim law, and in Germanic and Celtic customary law. The Germanic traditions were imported into England, where they combined with Norman concepts to become the basis of the English common law of marriage. The Germanic law provided, at least in higher class families with property, for a payment by the bridegroom for the transfer of the *Mund* (responsibility for, and power over, the woman); for a settlement on the groom by the bride's family; and for a so-called *Morgengab,* which may have represented the completion of the settlement. The giving of a ring had a symbolic role in many kinds of wedding and betrothal ceremonies. The word wed derives from the Anglo-Saxon word for security given to bind a promise. The property used as security was not necessarily transferred but given symbolically (*i.e.,* the ring). In a modern Church of England wedding service, the giving of security is reflected in the words "With this ring I thee wed," and the settlement of property in the words "and with all my worldly goods I thee endow." The minister has previously asked, "Who giveth this woman to be married to this man?" and, on "receiving the woman at her father's or friend's hands," proceeds with the ceremony. This "giving away" of the woman by her family reflects the transfer of the *Mund* to the bridegroom. In some systems the marriage forms may have a "bride purchase" origin, in the sense of compensation to her family (although there are differences of opinion as to the meaning of the customary forms); this was true in certain kinds of marriage in the earlier Roman republic, in Babylonian or Aramaic marriages, in early Arabic marriages, in certain Chinese unions (at least with regard to concubines, in which cases the transaction was more openly a purchase from the girl's parents), in customary marriage in some parts of Africa (*e.g.,* Nigeria, Ghana, Kenya), and in customary marriage among the nomadic tribes of Siberia (*e.g.,* the Kirgiz or Yakuts).

The ancient concept of marriage in many legal systems is that of a transaction between families (and this has sometimes persisted to the present day). Although the consent of the bride and bridegroom was almost always formally required, it may be questioned how real the consent was in the case of a child bride or in marriages between parties who did not see each other beforehand. Go-betweens and marriage brokers have been part of the marriage customs of many countries, especially in the East. The go-between and the professional marriage broker still have a following in some countries. Some have suggested that computer "matching" for marriage may become a modern development of the go-between concept.

Modern marriage law

Marriage as a voluntary relationship. The modern idea of marriage, which is becoming almost universal, is a voluntary exchange of promises between the man and the woman. Even though a marriage may involve substantial decisions as to property, these matters now tend either to be automatic (when there is no marriage contract) or to be formalized separately from the marriage ceremony. The ceremony itself is normally an exchange of consents accompanied by religious observances or a civil ceremony (or both). The purpose of the legal formalities is to differentiate the relationship from concubinage and to create certain recognized legal incidents such as maintenance, custody of children, rights under matrimonial regimes, intestate succession, claims under life-insurance policies and pension funds.

Legal limitations on marriage. In earlier legal systems, especially in Asia, the woman's consent was often unnecessary or of minor importance; the marriage negotiations took place between the woman's father and the man or his family. Voluntary consent of the parties became important in Roman times. Roman law during the period of the empire distinguished between an agreement for present marriage and an agreement for future marriage (*sponsalia per verba de praesenti* and *sponsalia per verba de futuro*). This distinction was taken over by Christianity, and a promise for marriage *per verba de futuro* was supported by a guarantee or "deposit" payment or by a penalty clause in a marriage contract.

Engagement. The view of the canon law of Christianity was that an engagement incapacitated a person from marriage to a different party and consequently provided ground for annulment of a marriage. This raised an issue that has troubled the civil lawyer but apparently not the common lawyer; *i.e.,* whether penalties, forfeiture provisions, damages, and the like for breach of engagement or betrothal are consistent with the exchange of voluntary consent at the marriage ceremony. Thus, French law has been led to reject an action of breach of promise (while permitting an action in delict—that is, on the ground that one party has been wronged). The common law, on the other hand, allows claims for breach of promise, although the modern tendency is to eliminate this form of action by statute.

The public interest. It has been difficult to delineate the boundaries between public and private interest in marriage law. The public interest is involved in the prevention of clandestine marriages; in requiring a license or the publication of banns as a condition precedent to marriage; in requiring parental consent for marriages between persons of certain ages; and in providing for the registration of marriages in a public manner. In practice, however, the marriage laws are often a mixture of functional administrative provisions (such as the requirement for registration and health certificates), old customs, and religious ceremonies. Marriage statutes were introduced in modern times to combat the danger of clandestine marriages, which were possible under the old law in Europe and England by some form of mutual consent. In addition to direct proof of consent, a clandestine marriage could be established by engagement followed by intercourse (*matrimonium subsequente copula*) or by habit and repute marriage (evidence of acceptance in the community as being married persons). Clandestine marriage was significant at a time when a man could acquire control over the property of a woman, including absolute ownership of much of it. The emancipation of women has put an end to the economic advantages of the clandestine marriage, but the legislation to which it gave rise has left an impress on the statute books.

Clandestine marriage

Age. In order to satisfy the requirement of a voluntary consent to a marriage, a party must have reached an age at which he or she is able to give a meaningful consent, and it is also implied that a person may be legally disqualified on mental grounds from having capacity to marry. Marriages of young children, negotiated by their parents, are prohibited in modern societies. Historically, the attitude of the English common law was that a person under seven years lacked the mental ability to consent to marriage, and that between seven years and puberty there could be consent but not a consummated marriage. At common law, therefore, the marriage of a person between the ages of seven and 12 or 14 was "inchoate" and would become "choate" on reaching puberty, if no objection was raised. Most modern legal systems provide for a legal minimum age of marriage ranging from 15 to 18 years. Some systems require parental consent to marriage when the parties are above the minimum age but below some other age, and failure to obtain this may be a ground for annulment. Parental consent has a long historical tradition, and there have been systems in which the girl's consent was virtually unnecessary. It is difficult to say, therefore, whether modern provisions have a valid social function or are the flotsam of older ideas on marriage.

Relationship. Other laws forbid marriage between persons having certain ties of relationship, either of blood or of marriage. "Forbidden degrees" of one sort or another exist in most social groups. The rules against marrying close relatives are sometimes said to be directed against the dangers of inbreeding, but this does not explain the prohibition against unions between persons who are related only by marriage. In classical Chinese society, marriage was regarded as a linking of different families, and the traditional pattern was exogamy (marriage outside the family). In ancient Egypt, on the other hand, where the pharaoh was deified, marriages within the blood were considered desirable in order to preserve its purity. Marriages between cousins are apparently encouraged in some Arab countries, perhaps to strengthen family ties and to keep the property together.

Religion. Religion has had a strong influence on marriage law, often providing the main basis of authority. Hindu family law, which goes back at least 4,000 years (and may be the oldest known system), is a branch of dharma—that is, the aggregate of religious, moral, social, and legal duties and obligations as developed by the

*Smṛti*s, or collections of the law. Islāmic and Jewish family law also rests on spiritual authority. Religious courts have had jurisdiction over family matters in various countries, and in some countries they still possess it. Some modern religious courts retain only their spiritual jurisdiction over marriage and divorce; their judgments have no standing in the secular law. In some Roman Catholic and Greek Orthodox Christian marriages and also in Muslim and Jewish marriages, the application of the religious law is regarded as binding upon persons belonging to the faith. Where religious texts provide the literal authority for legal principles, as in Islāmic law, it may be necessary to reinterpret the texts in order to reform the law. This raises complex issues in those Muslim countries where there are movements for greater equality of the sexes.

ECONOMIC ASPECTS OF FAMILY LAW

The variety of approaches to marital property

The property of husband and wife. The comparative legal history of marital property, viewed in broad perspective, consists of a period stretching back for about 4,000 years, during which a husband was generally regarded as a quasi-guardian of his wife, who was dependent upon him economically and legally. The English common law removed the separate legal personality of a woman when she married and merged it in that of her husband, though she regained it if she became a widow. Her husband acquired extensive rights to the administration and ownership of her property, including full ownership (with no obligation even to give an accounting) of any moneys she received from employment or business. Until quite recently, the only property of which a Hindu woman was the absolute owner was her *strīdhana,* consisting mainly of wedding gifts and gifts from relatives. Muslim women have traditionally owned and managed their own property. In Japan, before enactment of the Meiji Civil Code of 1898, all of the woman's property such as land or money passed to her husband except for personal clothing and a mirror stand. In early Vietnamese customary law, the property contributions of both spouses formed a common mass that was administered jointly and then divided, together with acquisitions made during their marriage, between the spouses and their heirs. In ancient Burmese law, all of the property of the spouses comprised a common mass that was their joint property but was administered by the husband, who exercised extensive powers. In 13th-century France there was a concept of "community" on the lines of a partnership between the spouses; but by the 16th century, French husbands had secured the *puissance maritale* ("marital power") that gave them power over the disposal of property. In ancient Hungarian law, the bride usually brought property to the marriage, and a regime with a form of dowry persisted until World War II. The law of imperial Russia permitted a wife to own and deal with property independently of her husband—an attitude that differed from its treatment of women in other respects. A Roman wife, in some marriage forms, had a position of independence in regard to her property that was unusual for that era. In Celtic law, the man gave property to the woman's family; according to Brehon law (the ancient law of Ireland), the payment could be made by annual installments over 21 years (the woman's father kept the first installment, two-thirds of the second, half of the third, and so on, the remainder going to the wife).

Brehon law

The emancipation of women, which occurred in many countries during the latter part of the 19th century and the first part of the 20th, had a profound effect upon family law and marital property. The Scandinavian countries made radical reforms in their marital property laws in the 1920s; they introduced a new type of matrimonial regime in which the spouses retain independent control of their property except for some items for the disposal of which the consent of the other spouse is required, but the combined remaining matrimonial property is divided at the termination of the regime. This has been influential in the reforms of other countries. The Federal Republic of Germany introduced legal equality of the sexes as a constitutional principle in the 1950s, followed by substantial amendments to the civil code with respect to the financial and property relations of spouses; these included a new matrimonial regime on sharing the value of the economic gains of marriage, analogous in broad principle to the Scandinavian reforms but containing new features. The French civil code received major revisions with respect to the matrimonial regimes in 1965. In 1969 the civil code of Quebec was revised to give full capacity to married women; it also created a new matrimonial regime, described as a partnership of acquests (that is, property acquired by the spouses after their marriage). In 1950 the People's Republic of China enacted a comprehensive marriage law including provisions giving the spouses equal rights with regard to ownership and management of marital property. The Soviet Union issued the Fundamentals of Legislation on Marriage and the Family in 1968, and a marriage and family code, providing for a form of community of acquests, was enacted in 1969 for the Russian S.F.S.R. The German Democratic Republic introduced a code of family law in 1965 based on principles of equality of the sexes in financial and property matters. Czechoslovakia's code of family law, enacted in 1964, introduced a concept of joint ownership with an equal share in management. Community sharing systems have been introduced by codes in Romania (1954) and Poland (1964). Substantial changes in the property capacity of Hindu women were made by the Hindu Succession Act of 1956.

In the 1970s the reform of the treatment at law of marital property became an important issue in various other countries, including England and Scotland, Belgium, Israel, and Canada. There was a quantity of new legislation on matrimonial property, especially in the common-law countries, and in some cases the process of reform and legislative change has continued.

Maintenance and support. The law of maintenance and support has differed from that of marital property in most countries. A widow, for example, normally receives some share in her husband's estate upon his death. Some systems of law require that dependents receive a compulsory share in the estate or dependent's relief or family provision (that is, financial support out of the estate for a dependent in straitened circumstances). Most systems of law have traditionally regarded financial support as the responsibility of a husband and a father (to which any dowry or property from the wife was considered a contribution). But in a hierarchical structure, such as the Greek or Roman family, the responsibility might rest on the paterfamilias, if different from the husband. In Hindu law, the male members of a joint family, together with their wives, widows, and children, are entitled to support out of the joint property.

Traditional laws of support

The influence of legislation. Social welfare legislation and the principle that a child's welfare is paramount have added a dimension and an inconsistency to the traditional principle of paternal responsibility. The new dimension is a public one and implies that society has an ultimate responsibility to see that children receive at least a minimum standard of maintenance. In some countries—for example, the United States, Canada, and various European countries—attempts have been made to combine parental and public responsibility for the child's welfare.

The enforcement of the legal obligation of a parent to maintain a child runs into a number of difficulties in law and practice. The father may be too poor to support his child, or he may be impossible to locate, or he may already be in prison for refusal to pay. A wife may be reluctant to sue her husband (for personal reasons and perhaps in the hope of a reconciliation), and if she does not do so, the child will be deprived of financial aid. Where there are social welfare programs supported by taxes, efforts may be made to protect the tax revenues by, for example, requiring a deserted wife to sue her husband as a condition of receiving welfare payments. Sometimes the authorities institute criminal or contempt proceedings against the husband. But in the North American and British experience, the payment record of obligants on maintenance orders is generally poor; the money recovered is usually inadequate; and the cost of enforcement may even approach the sums collected. It is not difficult to see why a high proportion of one-parent families exist in poverty even in countries with highly developed economies.

Working wives. The introduction to the British Census of 1851 stated: "The duties of a wife, a mother, a mistress of a family, can only be efficiently performed by unremitting attention; accordingly it is found that in districts where the women are much employed from home, the children and parents perish in great numbers." Even assuming that were true, much has changed since then. Most modern countries provide that a husband and wife have mutual obligations of financial support according to their means, aiming at a common standard of living for them both. The older principle was that the husband supported the wife in return for obtaining most of her property or becoming *chef de la famille et de la communauté* in a community-property system. Some common-law systems are a curious mixture of old and new principles. A high proportion of married women, in Western countries at least, are now employed outside the home, and the trend is expected to continue and probably accelerate. Two results of this trend are that (1) more married women now own some money and property and (2) on separation, divorce, or widowhood, a woman is more likely to be employed and so not dependent on support.

Separation of marital property. Reforms in marital-property laws have tended to reflect the wishes of spouses and their families, rather than traditional customs, religious attitudes, and dogmatic formulas. The French civil code of 1804 began a European pattern of giving spouses a choice of matrimonial regime: the codifiers were confronted with a variety of customary laws in different parts of the country, and, not wishing to impose one of them, they included alternatives in the code, designating one, the Custom of Paris, as the legal regime that would apply if the parties did not select another in a marriage contract. In common-law countries, the tendency has been to favour separation of property—a tendency resulting more by accident than by intention. This has come about because most of these countries adopted married women's property legislation that removed the incapacity of a married woman to make contracts and deal with her property, thus destroying the existing system by which the wife's property passed into the control of the husband. No new matrimonial system was constructed, so that the spouses were placed in the position of separate individuals so far as property was concerned. They can, of course, draw up marriage contracts or settlements to express their own wishes, but such contracts are now rare. The laws of the Communist countries usually provide a matrimonial regime, intended to promote the social and political policy of the government.

Alternatives in marital property law

In countries where the spouses are given a choice of regime, many of them do not exercise their option. A survey made in France in 1963–64 indicated that 76 percent of the couples made no marriage contract and therefore were subject to the legal regime. In Quebec, on the other hand, before the legislation in 1969 on marital property, more than 70 percent of marrying couples signed contracts for separation of property; part of the explanation for this may be that Quebec is surrounded by jurisdictions in which the common law prevails.

Even in modern times, in most cases husbands and wives differ in their potential for acquiring property. In separation of property, husbands and wives owning property and dealing with each other will be in the same position as unmarried adults. There are, however, grounds for distinguishing marital property questions from ordinary property questions, because persons who cohabit on a domestic basis share a common standard of living and usually also the benefits of each other's property. A major element in many marriages is the raising of children, and the traditional female role, requiring her full-time presence in the home, places the married woman at a disadvantage so far as earning money and acquiring property are concerned. It is inconsistent of society to encourage a woman to take the domestic role of wife and mother, with its lower money and property potential, but in property matters to treat her as if she were a single person. It is also inconsistent to place upon the husband the sole responsibility for maintaining his wife and children, if his wife has regular employment outside the home. When the marriage is dissolved, if the wife has not been regularly employed and now enters the labour market on a full-time basis, she may be at a considerable disadvantage as far as salary and pension rights are concerned.

Inconsistent social attitudes

Community property. A marital-property system should try to balance two sets of interests: the interests of the spouses and the interests of third parties such as purchasers, creditors, and business partners. Community-property regimes emphasize the first but are less attractive in terms of the second, because the property is tied up in the community and is subject to the interests of both spouses, whereas the third party may be dealing with only one of them. Separation of property gives property independence to each spouse, but it does not provide for sharing unless the spouses place items of property under joint ownership. Consequently, there has been a trend in many countries toward new regimes giving the husband and wife independence in dealing with property but also providing rules for a division of net assets on liquidation of the marriage. The first country in Europe to initiate this type was Sweden, in 1920.

In some common-law countries, certain items of property have received special treatment. Examples are the matrimonial home in English law (under decisions of the courts beginning around 1950) where there has been concern about a wife's being evicted by a purchaser or mortgagee from her husband, and also about division of proceeds of the sale when both spouses have contributed something to the purchase, or where one has done improvements on the other's property, or where the husband's payments toward the purchase have been possible because the wife's money was used for the day-to-day domestic needs of the family, and so on. In parts of the United States and Canada, there have been "homestead" laws providing that one spouse cannot dispose of or encumber the home without the consent of the other, that the use goes to the widow or family on the owner's death, and that the home cannot be sold in execution.

Tort actions between spouses. In English common law, as amended by the property legislation of the 19th century, a husband could not sue his wife in tort (that is, for a wrongful civil act not arising from contract), and she could sue him only in respect of damage to her separate property. This has been variously explained as stemming from the doctrine of the unity of the legal personalities of husband and wife (so that the plaintiff and the defendant are the same legal person) or from the belief that it would be disruptive to the family to allow damage suits between spouses. The modern tendency is to permit delict or tort action between spouses. This seems consistent with the fact that many damage suits, such as automobile accident claims, are covered by insurance, and the litigation in such cases is therefore between two insurance companies with the spouses as nominal parties.

Movements exist in North America and Europe favouring the recognition of a "no-fault" basis for certain delict or tort proceedings; this would transfer the emphasis in such actions to securing compensation for the person who suffered the damage, rather than determining whether the plaintiff can establish a cause of action (which usually means proving fault).

Co-ownership. Some marital-property systems that are basically separation of property have modifications for the situation in which, for example, an asset has been acquired by contributions from both spouses with the intention that both will benefit from its purchase—as with a home, furnishings, an automobile, a joint bank account, or joint investments. But the attitudes of husbands and wives as to their property after a marriage has broken down may be quite different from their intentions when an asset was acquired. There are decisions of the English courts that imply that in some of these circumstances, at least, the net value of the asset should be divided equally on the maxim that "equality is equity." The boundaries of this principle, however, are not at all certain.

Japanese marital-property law was revised in 1947, and the present legal regime is a modified form of separation of property. Under this regime, property to which only one spouse has title, but in the acquisition of which both have really cooperated during their marriage, is considered substantially co-owned. The civil code has been interpreted to the effect that substantially co-owned property is attributed to the title holder in a question involving third parties and to both spouses in a question between the spouses themselves.

DIVORCE

A marriage can terminate as a human relationship before it is dissolved by law. Quite often the court rulings as to property and the custody of children will merely confirm arrangements that have already been made by the parties concerned. In the United States and Canada, 80 to 90 percent of divorce proceedings are undefended; often the parties have made provisional arrangements about property, and one of them already has custody of the children. Before a union can be dissolved by divorce, there must have been a valid marriage. If a marriage has been imperfectly constituted in law, it may be annulled; grounds for annulment include lack of capacity, no reality of consent by the parties, a vitiating defect in the marriage ceremony, or the subsequent discovery of a defect such as inability to consummate the marriage.

In old legal systems, marriage was conceived as the transfer of a woman from the power of her family to that of her husband under terms usually specified in the marriage contract. The standard method of dissolving a marriage if both parties were alive was repudiation, resulting usually in the return of the woman to the power of her family. Repudiation has had a considerable history; it has strongly influenced marriage law in Muslim, Jewish, Chinese, and Japanese law. In Muslim law, repudiation can occur without proof of legally designated fault or a breakdown of the marriage. In practice, of course, there are checks on the too facile use of this power by a husband, such as objection from the wife's family, the obligation to repay the value of a dowry, or religious disapproval. In Roman marriage law, unilateral repudiation at will was permitted, and this freedom existed for some time in the early Christian Era. The concern of the Roman law was for solemnity rather than grounds, and unilateral divorce was by a notification of repudiation before seven witnesses.

Indissoluble marriages

At the other extreme from repudiation at will is the sacramental view of marriage (as in the teaching of certain Christian churches) that a marriage may not be dissolved during the joint lives of the spouses. Formerly, a Hindu marriage was indissoluble (except by caste custom) and might be eternal.

Between the extremes of repudiation at will and indissoluble marriage, there are various divorce formulas: divorce for fault, such as adultery, desertion, cruelty, or imprisonment; divorce on grounds analogous to frustration of contract, such as incurable insanity subsequent to the marriage or disappearance of the spouse; divorce by mutual agreement; and divorce on the ground that the marriage has broken down. The variety of laws and attitudes with respect to divorce is bewildering. In comparing divorce laws, it should be remembered that the rates of divorce and of marriage breakdown are different; the latter are hard to ascertain, although the breakdown rate obviously exceeds the divorce rate. If the grounds for divorce are widened, then the divorce rate will rise; if marriage is made indissoluble, the divorce rate will, of course, be zero. In many countries the trend has been toward more liberal divorce legislation (with a consequent rise in divorce rates).

A complicating factor in divorce law is the question of giving recognition to foreign divorces. The divorce laws of countries and states differ, and so do their rules for recognition of divorces elsewhere. A person living in a jurisdiction in which divorce is difficult to obtain may be able to go to another in which divorce laws are more liberal and obtain a dissolution of the marriage that will be recognized in the first jurisdiction. A feature of private international family law is the "limping" relationship—when a person is regarded as married by country X and as single by country Y, and a child as legitimate by country A and as illegitimate by country B. One reason why a country may restrict the recognition of divorces is that there are a number of jurisdictions in which divorce is granted on liberal grounds and with only nominal connections between the spouses and the divorce-granting jurisdiction (sometimes giving the impression of "divorce mills" that are operated for commercial reasons).

Some divorce laws provide for conciliation efforts (as in Sweden, Australia, and Canada), but these do not seem to have had any significant effect on divorce rates. In Chinese law, there is a long tradition of conciliation in many legal areas, including family law. But divorce stems from the desire to end an intimate human relationship that may have existed for some years. It is not an ordinary dispute at law; it has little in common with the interpretation of a business deal, a tax claim, a criminal charge, or other legal questions that can be presented fairly precisely. In a divorce, only the spouses can really know the differences between them, and neglect of this distinction can produce reasoning by false analogy.

FAMILY COURTS

In some countries there are special courts for family matters, set up in pursuit of religious, political, or social objectives; these include the Christian, Muslim, and Jewish ecclesiastical courts. There are also people's courts and conciliation courts, particularly in Communist countries.

Another approach has been to establish social courts that have a functional relation to the legal problems affecting families. Such problems include marriage, divorce, annulment, matrimonial regime, maintenance of spouses or of children, adoption, custody of children, legitimacy, filiation proceedings, juvenile delinquency, care and protection of children, assault on a spouse or a child, torts between spouses, marriage contracts, and judicial separation. Although these are the problems that produce the largest volume of private law litigation in most countries, family law has not, in many countries, been given a corresponding priority by the regular courts.

Character of family law

Those who favour special courts for family matters argue that family law is concerned with human relationships that require a judicial environment different from that of ordinary civil actions. The facts of the dispute in a family matter may not be as significant as the underlying problems (financial difficulties, health, addiction to drugs or alcohol) that have projected the issue. Another argument favouring family courts is that a high proportion of family proceedings are noncontentious or undefended; for example, proceedings concerning adoption and children in need of care normally require not so much the application of law as an inquiry into what is in the best interests of the child. In family matters, moreover, the court has need of ancillary services—social workers, probation officers, liaison with various social agencies. Since children and young people are often involved, there is need of special legal officers to present inquiry material to the court or to represent the interests of the children (which may conflict with the positions taken by their parents).

A number of countries have established special courts for cases relating to children and young people (sometimes with lay members) and special procedures for the disposition of such cases. Less progress has been made in the area of comprehensive family courts. One reason may be that family law can be less rewarding and more time-consuming as compared with more lucrative and prestigious fields of law.

It is sometimes argued that judges should be given a wide discretion in family cases. This seems to have been done in the family codes of eastern Europe and the People's Republic of China, which are loosely constructed with statements of politico-legal principles and leave much leeway to the judges or conciliators. In the United Kingdom and the United States the proposal has been made that in a divorce case, for example, the court should have discretion not only as to the custody of children but also as to arrangements for maintenance and disposition of property; the court should do what it thinks just, having regard to the history of the marriage and the behaviour of the spouses. Against this it is argued that such discretion would tend to turn tribunals into courts of morals in which the personal views of judges could prevail in the absence of applicable legal rules.

BIBLIOGRAPHY. Family law encompasses an enormous literature. Further, the principal legal writings in each country are in the language of that country, and there has been little translation of law books or articles. The following are a few suggestions for further reading, including a number of titles relating to the complex subject of comparative marital property.

Comparative law is the focus of MARY ANN GLENDON, *State, Law, and Family: Family Law in Transition in the United States and Western Europe* (1977), and a successor work, *The Transformation of Family Law: State, Law, and Family in the United States and Western Europe* (1989); JOHN EEKELAAR and SANFORD N. KATZ (eds.), *Marriage and Cohabitation in Contemporary Societies: Areas of Legal, Social, and Ethical Change* (1980); M.T. MEULDERS-KLEIN and JOHN EEKELAAR (eds.), *Family, State, and Individual Economic Security,* 2 vol. (1988), in English and French; JEAN PATARIN and IMRE ZAJTAY (eds.), *Le Régime matrimonial légal dans les législations contemporaines,* 2nd ed. (1974); and the journal *Revue internationale de droit comparé* (quarterly).

Works dealing with civil-law systems include VIRGILIO DE SÁ PEREIRA, *Direito de familia,* 2nd ed. (1959); MARCEL BRAZIER, *Le Nouveau Droit des époux et les régimes matrimoniaux* (1965); F.H. LAWSON, A.E. ANTON, and LIONEL NEVILLE BROWN (eds.), *Amos and Walton's Introduction to French Law,* 3rd ed. (1967); JEAN PATARIN and GEORGES MORIN, *La Réforme des régimes matrimoniaux,* 4th ed., vol. 1, *Statut fondamental et régime légal* (1977); JOACHIM GERNHUBER, *Lehrbuch des Familienrechts,* 3rd rev. ed. (1980); ANDRÉ COLOMER, *Droit civil: régimes matrimoniaux* (1982); GÜNTHER BEITZKE, *Familienrecht,* 25th ed. (1988); and JOSÉ CASTÁN TOBEÑAS, *Derecho civil español, común y foral,* vol. 5, *Derecho de familia,* 10th ed. rev. and updated by GABRIEL GARCÍA CANTERO and JOSÉ MA. CASTÁN VÁZQUEZ (1995).

General studies of common-law systems are RONALD H. GRAVESON and FRANCIS R. CRANE (eds.), *A Century of Family Law, 1857–1957* (1957); MARY ANN GLENDON, *The New Family and the New Property* (1981); JOHN EEKELAAR, *Family Law and Social Policy,* 2nd ed. (1984); MICHAEL D.A. FREEMAN (ed.), *The State, the Law, and the Family: Critical Perspectives* (1984); HOMER H. CLARK, JR., *The Law of Domestic Relations in the United States,* 2nd ed., 2 vol. (1987); LAURENCE D. HOULGATE, *Family and State: The Philosophy of Family Law* (1988); STEPHEN M. CRETNEY and J.M. MASSON, *Principles of Family Law,* 5th ed. (1990); HARRY D. KRAUSE (ed.), *Family Law,* 2 vol. (1992); and P.M. BROMLEY and N.V. LOWE, *Family Law,* 8th ed. (1992).

Works treating the subject of divorce and laws affecting the distribution of marital property include IAN F.G. BAXTER, *Marital Property* (1973); JUDITH S. WALLERSTEIN and JOAN BERLIN KELLY, *Surviving the Breakup: How Children and Parents Cope with Divorce* (1980); W.S. MCCLANAHAN, *Community Property Law in the United States* (1982); LENORE J. WEITZMAN, *The Divorce Revolution: The Unexpected Social and Economic Consequences for Women and Children in America* (1985); HERBERT JACOB, *Silent Revolution: The Transformation of Divorce Law in the United States* (1988); JUDITH S. WALLERSTEIN and SANDRA BLAKESLEE, *Second Chances: Men, Women, and Children a Decade After Divorce* (1989); and STEPHEN D. SUGARMAN and HERMA HILL KAY (eds.), *Divorce Reform at the Crossroads* (1990).

Discussions of child custody and child welfare can be found in JOSEPH GOLDSTEIN, ANNA FREUD, and ALBERT J. SOLNIT, *Beyond the Best Interests of the Child,* new ed. (1979); JEFF ATKINSON, *Modern Child Custody Practice,* 2 vol. (1986); SAMUEL M. DAVIS and MORTIMER D. SCHWARTZ, *Children's Rights and the Law* (1987); ANDREW BAINHAM and STEPHEN M. CRETNEY, *Children: The Modern Law* (1993); and DONALD T. KRAMER (ed.), *Legal Rights of Children,* 2nd ed., 3 vol. (1994).

Reproductive issues and surrogacy are examined in SHEILA MCLEAN (ed.), *Legal Issues in Human Reproduction* (1989); ELAINE SUTHERLAND and ALEXANDER MCCALL SMITH (eds.), *Family Rights: Family Law and Medical Advance* (1990), discussing the impact of rapid scientific change on the law; LARRY GOSTIN (ed.), *Surrogate Motherhood: Politics and Privacy* (1990); and MARTHA A. FIELD, *Surrogate Motherhood,* expanded ed. (1990).

Studies of eastern European systems of family law are KAZIMIERZ GRZYBOWSKI, *Soviet Legal Institutions* (1962, reprinted 1982); EDWARD L. JOHNSON, *An Introduction to the Soviet Legal System* (1969); DOMINIK LASOK, *Polish Family Law* (1968); ENDRE NIZSALOVSZKY, *Order of the Family,* trans. from Hungarian (1968); and ŠTEFAN LUBY (ed.), *Le Droit civil tchécoslovaque* (1969).

Family law in Muslim society is explored in JOSEPH SCHACHT, *An Introduction to Islamic Law* (1964, reprinted 1982); YVON LINANT DE BELLEFONDS, *Traité de droit musulman comparé,* 3 vol. (1965–73); ASAF A.A. FYZEE, *Outlines of Muhammadan Law,* 4th ed. (1974); B.R. VERMA, *Islamic Law—Personal,* 6th ed. rev. by M.H. BEG and S.K. VERMA (1986); C.G. WEERAMANTRY, *Islamic Jurisprudence: An International Perspective* (1988); and MOHAMMAD HASHIM KAMALI, *Principles of Islamic Jurisprudence,* rev. ed. (1991).

Hindu law is discussed in J. DUNCAN M. DERRETT, *Essays in Classical and Modern Hindu Law,* 4 vol. (1976–78); and DINSHAH H. MULLA, *Principles of Hindu Law,* 15th ed. by SUNDERLAL T. DESAI (1982).

Assessments of family law in Asian and African systems are presented in ROBERT LINGAT, *Les Régimes matrimoniaux du sud-est de l'Asie,* 2 vol. (1952–55); ARTHUR TAYLOR VON MEHREN (ed.), *Law in Japan* (1963); J.N.D. ANDERSON (ed.), *Family Law in Asia and Africa* (1968); DAVID C. BUXBAUM (ed.), *Family Law and Customary Law in Asia* (1968), and *Chinese Family Law and Social Change in Historical and Comparative Perspective* (1978); B.P. BERI, *Law of Marriage and Divorce in India* (1982); and HIROSHI ODA, *Japanese Law* (1992).

(I.F.G.B./Ed.)

Faraday

Michael Faraday, who became one of the greatest scientists of the 19th century, began his career as a chemist. He wrote a manual of practical chemistry that reveals his mastery of the technical aspects of his art, discovered a number of new organic compounds, among them benzene, and was the first to liquefy a "permanent" gas (*i.e.*, one that was believed to be incapable of liquefaction). His major contribution, however, was in the field of electricity and magnetism. He was the first to produce an electric current from a magnetic field, invented the first electric motor and dynamo, demonstrated the relation between electricity and chemical bonding, discovered the effect of magnetism on light, and discovered and named diamagnetism, the peculiar behaviour of certain substances in strong magnetic fields. He provided the experimental, and a good deal of the theoretical, foundation upon which James Clerk Maxwell erected classical electromagnetic field theory.

By courtesy of the National Portrait Gallery, London

Faraday, oil painting by T. Phillips, 1842. In the National Portrait Gallery, London.

Early life. Michael Faraday was born on September 22, 1791, in the country village of Newington, Surrey, now a part of South London. His father was a blacksmith who had migrated from the north of England earlier in 1791 to look for work. His mother was a country woman of great calm and wisdom who supported her son emotionally through a difficult childhood. Faraday was one of four children, all of whom were hard put to get enough to eat, since their father was often ill and incapable of working steadily. Faraday later recalled being given one loaf of bread that had to last him for a week. The family belonged to a small Christian sect, called Sandemanians, that provided spiritual sustenance to Faraday throughout his life. It was the single most important influence upon him and strongly affected the way in which he approached and interpreted nature.

Faraday received only the rudiments of an education, learning to read, write, and cipher in a church Sunday school. At an early age he began to earn money by delivering newspapers for a book dealer and bookbinder, and at the age of 14 he was apprenticed to the man. Unlike the other apprentices, Faraday took the opportunity to read some of the books brought in for rebinding. The article on electricity in the third edition of the *Encyclopædia Britannica* particularly fascinated him. Using old bottles and lumber, he made a crude electrostatic generator and did simple experiments. He also built a weak voltaic pile with which he performed experiments in electrochemistry.

Faraday's great opportunity came when he was offered a ticket to attend chemical lectures by Sir Humphry Davy at the Royal Institution of Great Britain in London. Faraday went, sat absorbed with it all, recorded the lectures in his notes, and returned to bookbinding with the seemingly unrealizable hope of entering the temple of science. He sent a bound copy of his notes to Davy along with a letter asking for employment, but there was no opening. Davy did not forget, however, and, when one of his laboratory assistants was dismissed for brawling, he offered Faraday a job. Faraday began as Davy's laboratory assistant and learned chemistry at the elbow of one of the greatest practitioners of the day. It has been said, with some truth, that Faraday was Davy's greatest discovery.

Studies with Sir Humphry Davy

When Faraday joined Davy in 1812, Davy was in the process of revolutionizing the chemistry of the day. Antoine-Laurent Lavoisier, the Frenchman generally credited with founding modern chemistry, had effected his rearrangement of chemical knowledge in the 1770s and 1780s by insisting upon a few simple principles. Among these was that oxygen was a unique element, in that it was the only supporter of combustion and was also the element that lay at the basis of all acids. Davy, after having discovered sodium and potassium by using a powerful current from a galvanic battery to decompose oxides of these elements, turned to the decomposition of muriatic (hydrochloric) acid, one of the strongest acids known. The products of the decomposition were hydrogen and a green gas that supported combustion and that, when combined with water, produced an acid. Davy concluded that this gas was an element, to which he gave the name chlorine, and that there was no oxygen whatsoever in muriatic acid. Acidity, therefore, was not the result of the presence of an acid-forming element but of some other condition. What else could that condition be but the physical form of the acid molecule itself? Davy suggested, then, that chemical properties were determined not by specific elements alone but also by the ways in which these elements were arranged in molecules. In arriving at this view he was influenced by an atomic theory that was also to have important consequences for Faraday's thought. This theory, proposed in the 18th century by Ruggero Giuseppe Boscovich, argued that atoms were mathematical points surrounded by alternating fields of attractive and repulsive forces. A true element comprised a single such point, and chemical elements were composed of a number of such points, about which the resultant force fields could be quite complicated. Molecules, in turn, were built up of these elements, and the chemical qualities of both elements and compounds were the results of the final patterns of force surrounding clumps of point atoms. One property of such atoms and molecules should be specifically noted: they can be placed under considerable strain, or tension, before the "bonds" holding them together are broken. These strains were to be central to Faraday's ideas about electricity.

Faraday's second apprenticeship, under Davy, came to an end in 1820. By then he had learned chemistry as thoroughly as anyone alive. He had also had ample opportunity to practice chemical analyses and laboratory techniques to the point of complete mastery, and he had developed his theoretical views to the point that they could guide him in his researches. There followed a series of discoveries that astonished the scientific world.

Faraday achieved his early renown as a chemist. His reputation as an analytical chemist led to his being called as an expert witness in legal trials and to the building up of a clientele whose fees helped to support the Royal Institution. In 1820 he produced the first known compounds of carbon and chlorine, C_2Cl_6 and C_2Cl_4. These compounds were produced by substituting chlorine for hydrogen in "olefiant gas" (ethylene), the first substitution reactions induced. (Such reactions later would serve to challenge the dominant theory of chemical combination proposed by Jöns Jacob Berzelius.) In 1825, as a result of research on illuminating gases, Faraday isolated and described benzene. In the 1820s he also conducted investigations of steel alloys, helping to lay the foundations for scientific metallurgy and metallography. While completing an assignment from the Royal Society of London to improve the quality of optical glass for telescopes, he produced a glass of very high refractive index that was to lead him, in 1845, to the discovery of diamagnetism. In 1821 he married Sarah Barnard, settled permanently at the Royal Institution, and began the series of researches on electricity and magnetism that was to revolutionize physics.

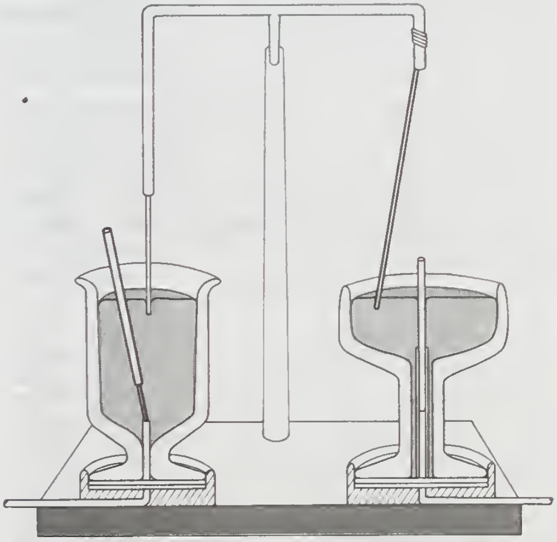
Discovery of electromagnetic rotation

In 1820 Hans Christian Ørsted had announced the discovery that the flow of an electric current through a wire produced a magnetic field around the wire. André-Marie Ampère showed that the magnetic force apparently was a circular one, producing in effect a cylinder of magnetism around the wire. No such circular force had ever before been observed, and Faraday was the first to understand what it implied. If a magnetic pole could be isolated, it ought to move constantly in a circle around a current-carrying wire. Faraday's ingenuity and laboratory skill enabled him to construct an apparatus that confirmed this conclusion (see Figure). This device, which transformed electrical energy into mechanical energy, was the first electric motor.

This discovery led Faraday to contemplate the nature of electricity. Unlike his contemporaries, he was not convinced that electricity was a material fluid that flowed through wires like water through a pipe. Instead, he thought of it as a vibration or force that was somehow transmitted as the result of tensions created in the conductor. One of his first experiments after his discovery of electromagnetic rotation was to pass a ray of polarized light through a solution in which electrochemical decomposition was taking place in order to detect the intermolecular strains that he thought must be produced by the passage of an electric current. During the 1820s he kept coming back to this idea, but always without result.

In the spring of 1831 Faraday began to work with Charles (later Sir Charles) Wheatstone on the theory of sound, another vibrational phenomenon. He was particularly fascinated by the patterns (known as Chladni figures) formed in light powder spread on iron plates when these plates were thrown into vibration by a violin bow. Here was demonstrated the ability of a dynamic cause to create a static effect, something he was convinced happened in a current-carrying wire. He was even more impressed by the fact that such patterns could be induced in one plate by bowing another nearby. Such acoustic induction is apparently what lay behind his most famous experiment. On August 29, 1831, Faraday wound a thick iron ring on one side with insulated wire that was connected to a battery. He then wound the opposite side with wire connected to a galvanometer. What he expected was that a "wave" would be produced when the battery circuit was closed and that the wave would show up as a deflection of the galvanometer in the second circuit. He closed the primary circuit and, to his delight and satisfaction, saw the galvanometer needle jump. A current had been induced in the secondary coil by one in the primary. When he opened the circuit, however, he was astonished to see the galvanometer jump in the opposite direction. Somehow, turning off the current also created an induced current in the secondary circuit, equal and opposite to the original current. This phenomenon led Faraday to propose what he called the "electrotonic" state of particles in the wire, which he considered a state of tension. A current thus appeared to be the setting up of such a state of tension or the collapse of such a state. Although he could not find experimental evidence for the electrotonic state, he never entirely abandoned the concept, and it shaped most of his later work.

Adapted from an illustration in *Dictionary of Scientific Biography*, edited by C C Gillespie, vol 4 copyright © 1971 American Council of Learned Societies, used with the permission of Charles Scribner's Sons



Faraday's apparatus for demonstrating the circular field of magnetic force around a current-carrying conductor. An electric current passes through two beakers of mercury, each containing a cylindrical bar magnet. One magnet (left) is freely pivoted, and the upper end rotates about the wire when the current is flowing. The other magnet (right) is fixed, and the pivoted wire rotates about it.

In the fall of 1831 Faraday attempted to determine just how an induced current was produced. His original experiment had involved a powerful electromagnet, created by the winding of the primary coil. He now tried to create a current by using a permanent magnet. He discovered that when a permanent magnet was moved in and out of a coil of wire a current was induced in the coil. Magnets, he knew, were surrounded by forces that could be made visible by the simple expedient of sprinkling iron filings on a card held over them. Faraday saw the "lines of force" thus revealed as lines of tension in the medium, namely air, surrounding the magnet, and he soon discovered the law determining the production of electric currents by magnets: the magnitude of the current was dependent upon the number of lines of force cut by the conductor in unit time. He immediately realized that a continuous current could be produced by rotating a copper disk between the poles of a powerful magnet and taking leads off the disk's rim and centre. The outside of the disk would cut more lines than would the inside, and there would thus be a continuous current produced in the circuit linking the rim to the centre. This was the first dynamo. It was also the direct ancestor of electric motors, for it was only necessary to reverse the situation, to feed an electric current to the disk, to make it rotate.

The first dynamo

While Faraday was performing these experiments and presenting them to the scientific world, doubts were raised about the identity of the different manifestations of electricity that had been studied. Were the electric "fluid" that apparently was released by electric eels and other electric fishes, that produced by a static electricity generator, that of the voltaic battery, and that of the new electromagnetic generator all the same? Or were they different fluids following different laws? Faraday was convinced that they were not fluids at all but forms of the same force, yet he recognized that this identity had never been satisfactorily shown by experiment. For this reason he began, in 1832, what promised to be a rather tedious attempt to prove that all electricities had precisely the same properties and caused precisely the same effects. The key effect was electrochemical decomposition. Voltaic and electromagnetic electricity posed no problems, but static electricity did. As Faraday delved deeper into the problem, he made two startling discoveries. First, electrical force did not, as had long been supposed, act at a distance upon chemical molecules to cause them to dissociate. It was the passage of electricity through a conducting liquid medium that caused the molecules to dissociate, even when the electricity merely discharged into the air and did not pass into a "pole" or "centre of action" in a voltaic cell. Second, the amount of the decomposition was found to be related in a simple manner to the amount of electricity that passed

Theory of electrochemistry

through the solution. These findings led Faraday to a new theory of electrochemistry. The electric force, he argued, threw the molecules of a solution into a state of tension (his electrotonic state). When the force was strong enough to distort the fields of forces that held the molecules together so as to permit the interaction of these fields with neighbouring particles, the tension was relieved by the migration of particles along the lines of tension, the different species of atoms migrating in opposite directions. The amount of electricity that passed, then, was clearly related to the chemical affinities of the substances in solution. These experiments led directly to Faraday's two laws of electrochemistry: (1) The amount of a substance deposited on each electrode of an electrolytic cell is directly proportional to the quantity of electricity passed through the cell. (2) The quantities of different elements deposited by a given amount of electricity are in the ratio of their chemical equivalent weights.

Faraday's work on electrochemistry provided him with an essential clue for the investigation of static electrical induction. Since the amount of electricity passed through the conducting medium of an electrolytic cell determined the amount of material deposited at the electrodes, why should not the amount of electricity induced in a nonconductor be dependent upon the material out of which it was made? In short, why should not every material have a specific inductive capacity? Every material does, and Faraday was the discoverer of this fact.

By 1839 Faraday was able to bring forth a new and general theory of electrical action. Electricity, whatever it was, caused tensions to be created in matter. When these tensions were rapidly relieved (*i.e.*, when bodies could not take much strain before "snapping" back), then what occurred was a rapid repetition of a cyclical buildup, breakdown, and buildup of tension that, like a wave, was passed along the substance. Such substances were called conductors. In electrochemical processes the rate of buildup and breakdown of the strain was proportional to the chemical affinities of the substances involved, but again the current was not a material flow but a wave pattern of tensions and their relief. Insulators were simply materials whose particles could take an extraordinary amount of strain before they snapped. Electrostatic charge in an isolated insulator was simply a measure of this accumulated strain. Thus, all electrical action was the result of forced strains in bodies.

The strain on Faraday of eight years of sustained experimental and theoretical work was too much, and in 1839 he suffered a breakdown of his health. For the next six years he did little creative science. Not until 1845 was he able to pick up the thread of his researches and extend his theoretical views.

Later life. Since the very beginning of his scientific work, Faraday had believed in what he called the unity of the forces of nature. By this he meant that all the forces of nature were but manifestations of a single universal force and ought, therefore, to be convertible into one another. In 1846 he made public some of the speculations to which this view led him. A lecturer, scheduled to deliver one of the Friday evening discourses at the Royal Institution by which Faraday encouraged the popularization of science, panicked at the last minute and ran out, leaving Faraday with a packed lecture hall and no lecturer. On the spur of the moment, Faraday offered "Thoughts on Ray Vibrations." Specifically referring to point atoms and their infinite fields of force, he suggested that the lines of electric and magnetic force associated with these atoms might, in fact, serve as the medium by which light waves were propagated. Many years later, Maxwell was to build his electromagnetic field theory upon this speculation.

When Faraday returned to active research in 1845, it was to tackle again a problem that had obsessed him for years, that of his hypothetical electrotonic state. He was still convinced that it must exist and that he simply had not yet discovered the means for detecting it. Once again he tried to find signs of intermolecular strain in substances through which electrical lines of force passed, but again with no success. It was at this time that a young Scot, William Thomson (later Lord Kelvin), wrote Faraday that he had studied Faraday's papers on electricity and magnetism and that he, too, was convinced that some kind of strain must exist. He suggested that Faraday experiment with magnetic lines of force, since these could be produced at much greater strengths than could electrostatic ones.

Faraday took the suggestion, passed a beam of plane-polarized light through the optical glass of high refractive index that he had developed in the 1820s, and then turned on an electromagnet so that its lines of force ran parallel to the light ray. This time he was rewarded with success. The plane of polarization was rotated, indicating a strain in the molecules of the glass. But, once again, Faraday noted an unexpected result. When he changed the direction of the ray of light, the rotation remained in the same direction, a fact that Faraday correctly interpreted as meaning that the strain was not, after all, in the molecules of the glass but in the magnetic lines of force. The direction of rotation of the plane of polarization depended solely upon the polarity of the lines of force; the glass served merely to detect the effect.

Discovery of diamagnetism

This discovery confirmed Faraday's faith in the unity of forces, and he plunged onward, certain that all matter must exhibit some response to a magnetic field. To his surprise he found that this was in fact so, but in a peculiar way. Some substances, such as iron, nickel, cobalt, and oxygen, lined up in a magnetic field so that the long axes of their crystalline or molecular structures were parallel to the lines of force; others lined up perpendicular to the lines of force. Substances of the first class moved toward more intense magnetic fields; those of the second moved toward regions of less magnetic force. Faraday named the first group paramagnetics and the second diamagnetics. After further research he concluded that paramagnetics were bodies that conducted magnetic lines of force better than did the surrounding medium, whereas diamagnetics conducted them less well. By 1850 Faraday had evolved a radically new view of space and force. Space was not "nothing," the mere location of bodies and forces, but a medium capable of supporting the strains of electric and magnetic forces. The energies of the world were not localized in the particles from which these forces arose but rather were to be found in the space surrounding them. Thus was born field theory. As Maxwell later freely admitted, the basic ideas for his mathematical theory of electrical and magnetic fields came from Faraday; his contribution was to mathematize those ideas in the form of his classical field equations.

From about 1855, Faraday's mind began to fail. He still did occasional experiments, one of which involved attempting to find an electrical effect of raising a heavy weight, since he felt that gravity, like magnetism, must be convertible into some other force, most likely electrical. This time he was disappointed in his expectations, and the Royal Society refused to publish his negative results. More and more, Faraday began to sink into senility. Queen Victoria rewarded his lifetime of devotion to science by granting him the use of a house at Hampton Court and even offered him the honour of a knighthood. Faraday gratefully accepted the cottage but rejected the knighthood; he would, he said, remain plain Mr. Faraday to the end. He died on August 25, 1867, and was buried in Highgate Cemetery, London, leaving as his monument a new conception of physical reality.

BIBLIOGRAPHY. An exhaustive account of Faraday's life and work is L. PEARCE WILLIAMS, *Michael Faraday* (1965, reissued 1971). Faraday's ideas on field theory and their later development by Maxwell are treated in L. PEARCE WILLIAMS, *The Origins of Field Theory* (1967, reissued 1981). Earlier biographies still worth consulting are JOHN TYNDALL, *Faraday As a Discoverer* (1869, reissued 1961); and SILVANUS P. THOMPSON, *Michael Faraday: His Life and Work* (1898). JOSEPH AGASSI, *Faraday As a Natural Philosopher* (1971), described as a historical novel, is interesting but untrustworthy as an account of Faraday's life and thought. Faraday's ideas can be found in his *Experimental Researches in Electricity*, 3 vol. (1839–55; reissued in 2 vol., 1965). *The Selected Correspondence of Michael Faraday*, ed. by L. PEARCE WILLIAMS, 2 vol. (1971), follows Faraday's ideas and discourses with colleagues on a host of subjects.

(L.P.W.)

Farming and Agricultural Technology

In spite of the often concentrated industrialization of many parts of the world, the ancient calling of agriculture continues to draft into its service more of the world's aggregate manpower than all other occupations combined. Farming can be described as the art of making land more productive, and agricultural technology can be described as the application of techniques to control the growth and harvesting of animal and vegetable products. This technology has a long history, dating back to the beginnings of the human race. While farming is still practiced in some areas by methods not far removed from the conditions of several thousand years ago, this article is concerned with relatively modern techniques, discussing modern agriculture as a practical art, a commercial enterprise, and a science.

Historical aspects of farming technologies described in this article are covered in the articles AGRICULTURE, THE HISTORY OF; and TECHNOLOGY, THE HISTORY OF. The academic disciplines that study, develop, and teach scientific techniques of farming are discussed in AGRICULTURAL SCIENCES. The processing and distribution of agricultural commodities are treated in FOOD PROCESSING and BEVERAGE PRODUCTION. FORESTRY AND WOOD PRODUCTION and GARDENING AND HORTICULTURE cover sciences that are closely related to farming, and the manufacture of synthetic fertilizers is covered in INDUSTRIES, CHEMICAL PROCESS. The changing role of agriculture in developing countries is taken up in ECONOMIC GROWTH AND PLANNING and LAND REFORM AND TENURE. For information on farming and agricultural technology in specific countries or regions, see articles such as CHINA; AFRICA; or UNITED STATES: *Wisconsin.*

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 533 and 731, and the *Index.* (Ed.)

This article is divided into the following sections:

AGRICULTURAL ECONOMICS

Agriculture is the source of livelihood for more than half of the world's population. In some countries more than four-fifths of the inhabitants support themselves by farming, while in the more industrialized countries the proportion ranges much lower—to less than 3 percent in both the United States and Great Britain. In general one can say that, when a large fraction of a nation's population depends on agriculture for its livelihood, average incomes are low. This does not mean that a nation is poor because most of its population is engaged in agriculture; it is closer to the truth to say that because a country is poor most of its people must rely upon agriculture for a living.

Agriculture and economic development

The rural surplus

As a country develops economically, the relative importance of agriculture declines. The primary reason for this was shown by the 19th-century German statistician Ernst Engel, who discovered that as incomes increase the proportion of income spent on food declines. For example, if a family's income were to increase by 100 percent, the amount it would spend on food might increase by 60 percent; if formerly its expenditures on food had been 50 percent of its budget, after the increase they would amount to only 40 percent of its budget. It follows from this that, as incomes increase, a smaller fraction of the total resources of society is required to produce the amount of food demanded by the population.

PROGRESS IN FARMING

This fact would have surprised most economists of the early 19th century, who feared that the limited supply of land in the populated areas of Europe would determine that continent's ability to feed its growing population. Their fear was based on the so-called law of diminishing returns: that under given conditions an increase in the amount of labour and capital applied to a fixed amount of land results in a less than proportional increase in the output of food. This principle is a valid one, but what the classical economists could not foresee was the extent to which the state of the arts and the methods of production would change. Some of the changes occurred in agriculture; others occurred in other sectors of the economy but had a major effect on the supply of food.

In looking back upon the history of the more developed countries, one can see that agriculture has played an important part in the process of their enrichment. For one thing, if development is to occur, agriculture must be able to produce a surplus of food to maintain the growing nonagricultural labour force. Since food is more essential for life than are the services provided by merchants or bankers or factories, an economy cannot shift to such activities unless food is available for barter or sale in sufficient quantities to support those engaged in them. Unless food can be obtained through international trade, a country does not normally develop industrially until its farm areas can supply its towns with food in exchange for the products of their factories.

Economic development also requires a growing labour force. In an agricultural country most of the workers needed must come from the rural population. Thus agriculture must not only supply a surplus of food for the towns, but it must also be able to produce the increased amount of food with a relatively smaller labour force. It may do so by substituting animal power for human power or by gradually introducing labour-saving machinery.

Agriculture may also be a source of the capital needed for industrial development to the extent that it provides a surplus that may be converted into the funds needed to purchase industrial equipment or to build roads and provide public services.

For these reasons a country seeking to develop its economy may be well advised to give a significant priority to agriculture. Experience in the developing countries has shown that agriculture can be made much more productive with the proper investment in irrigation systems, research, fertilizers, insecticides, and herbicides.

Fortunately, many advances in applied science do not require massive amounts of capital, although it may be necessary to expand marketing and transportation facilities so that farm output can be brought to the entire population.

One difficulty in giving priority to agriculture is that most of the increase in farm output and most of the income gains are concentrated in certain regions rather than extending throughout the country. The remaining farmers are not able to produce more and actually suffer a disadvantage as farm prices decline. There is no easy answer to this problem, but developing countries need to be aware of it; economic progress is consistent with lingering backwardness, as can be seen in parts of southern Italy or in the Appalachian area of the United States.

PEASANT AGRICULTURE

One characteristic of undeveloped peasant agriculture is its self-sufficiency. Farm families in those circumstances consume a substantial part of what they produce. While some of their output may be sold in the market, their total production is generally not much larger than what is needed for the maintenance of the family. Not only is productivity per worker low under these conditions but yields per unit of land are also low. Even where the land was originally fertile, the fertility is likely to have been depleted by decades of continuous cropping. The available manures are not sufficient, and the farmers cannot afford to purchase them elsewhere.

Peasant agriculture is often said to be characterized by inertia. The peasant farmer is likely to be illiterate, suspicious of outsiders, and reluctant to try new methods; food patterns remain unchanged for decades or even centuries. Evidence, however, suggests that the apparent inertia may be simply the result of a lack of alternatives. If there is nothing better to change to, there is little point in changing. Moreover, the self-sufficient farmer is bound to want to minimize his risks; since a crop failure can mean starvation in many parts of the world, farmers have been reluctant to adopt new methods if doing so would expose them to greater risks of failure.

The increased use worldwide of high-yielding varieties of rice and wheat since the 1960s has shown that farmers are willing and able to adopt new crops and farming methods when their superiority is demonstrated. These high-yielding varieties, however, require increased outlays for fertilizer, as well as expanded facilities for storage and distribution, and many developing countries are unable to afford such expenditures.

THE LABOUR FORCE

As economic development proceeds, a large proportion of the farm labour force must shift from agriculture into other pursuits. This fundamental shift in the labour force is made possible, of course, by an enormous increase in output per worker as agriculture becomes modernized. This increase in output stems from various factors. Where land is plentiful the output per worker is likely to be higher because it is possible to employ more fertilizer and machinery per worker.

Land, output, and yields

Only a small fraction of the world's land area—about one-tenth—may be considered arable, if arable land is defined as land planted to crops. Less than one-fourth of the world's land area is in permanent meadows and pastures. The remainder is either in forests or is not being used for agricultural purposes.

GENERAL RELATIONSHIPS

Land and population

There are great differences in the amount of arable land per person in the various regions of the world. The greatest amount of arable land per capita is in Oceania; the least is in China. No direct relationship exists between the amount of arable land per capita and the level of income; Europe has almost as little arable land per capita as Asia and less than Africa; Japan and The Netherlands have very limited amounts of arable land per capita.

The relationship between land, population, and farm production is a complex one. In traditional agriculture, where methods of production have changed little over a long period of time, production is largely determined by the quality and quantity of land available and the number of people working on the land. Until the early years of the 20th century, most of the world's increase in crop production came either from an increase in land under cultivation or from an increase in the amount of labour used per unit of land. This generally involved a shift to crops that would yield more per unit of land and required more labour for their cultivation. Wheat, rye, and millet require less labour per unit of land and per unit of food output than do rice, potatoes, or corn (maize), but generally the latter yield more food per unit of land. Thus, as population density increased, the latter groups of crops tended to be substituted for the former. This did not hold true in Europe, where wheat, rye, and millet expanded at the expense of pasture land; but these crops yielded more food per acre than did the livestock that they displaced.

Impact of modernization

As agriculture becomes modernized, its dependence upon land as well as upon human labour decreases. Animal power and machinery are substituted for human labour; mechanical power then replaces animal power. The substitution of mechanical power for animal power also reduces the need for land. The increased use of fertilizer as modernization occurs also acts as a substitute for both land and labour; the same is true of herbicides and insecticides. By making it possible to produce more per unit of land and per hour of work, less land and labour are required for a given amount of output.

RECENT TRENDS

Crop yields have increased dramatically since 1950, with a faster rate of growth in the developing than in the developed countries. Most of this increased output has been due to gains in yields rather than to the expansion of cultivated land.

In Europe as well as in North and Central America, the total area under crops has declined; in South America it has increased by more than one-half and in Asia by more than one-third. The large increase in Oceania was due to immigration. The large decrease in Africa was due to a succession of droughts from the 1970s on.

Grain yields in the developed regions of the world have increased consistently over the past several decades. In the rest of the world the pre-World War II yields were not achieved again until the mid-1950s. The increases in grain production were more than twice as high in the developing as in the developed countries.

Food production and total agricultural production exhibit nearly identical trends, and changes in food production can be taken therefore as indicative of changes in total agricultural production. Food supplies per capita in developing countries have increased at nearly the same rate as in developed countries, indicating a narrowing gap between food supplies and population growth in the developing countries.

Efforts to control prices and production

The need for control

In the past few decades governments have undertaken to control both prices and output in the agricultural sector, largely in response to the pressures of the farmers themselves. In the absence of such control, farm prices tend to fluctuate more than do most other prices, and the incomes of farmers fluctuate to an even greater degree. Not only are incomes in agriculture unstable, but they also tend to be lower than incomes in other economic sectors.

THE PROBLEM

Instability of prices. The instability of farm prices results from several factors. One is the relative slowness with which farmers are able to respond to changes in the demand for their product. Farmers generally must produce on the basis of expectations, and if their expectations turn out to be wrong, the resulting surplus or shortage cannot be corrected until the beginning of the next production cycle. Once a crop is planted, very little can be done to increase or decrease production in response to market prices. As long as prices cover current operating costs, such as the cost of harvesting, it pays farmers to carry through their production plans even if prices fall to a very low level. It is not unusual for the prices of particular farm products to vary by a third or a half from year to year. This extreme variability results from the relatively low responsiveness of demand to changes in price—*i.e.*, from the fact that in order to increase sales by 5 percent it may be necessary to reduce the price by 15 percent.

Instability of income. The instability of farm prices is accompanied by instability of farm income. While gross income from agriculture generally does not vary as much as do individual farm prices, net income may vary more than prices. In modern agriculture costs tend to be relatively stable; the farmer is unable to compensate for a drop in prices by reducing his payments for machinery, fertilizer, or labour.

Relative stability of costs and its effect on income

The incomes of farm workers are generally below those of other workers. There are two major reasons for this inequity. One is that in most economies the need for farm labour is declining, and each year large numbers of farm people, especially young ones, must leave their homes to seek jobs elsewhere. The difference in returns to labour is required to bring about this transfer of workers out of farming; if the transfer did not occur, farm incomes would be even more depressed. The second major reason for the income differences is that farm people generally have less education than do nonfarm people and are able to earn less at nonfarm jobs. The difference in education is of long standing and is found in all countries, developed and undeveloped; it also exists whether the national education system is highly decentralized, as in the United States, or highly centralized, as in France.

GOVERNMENT INTERVENTION

Governments have employed various measures to maintain farm prices and incomes above what the market would otherwise have yielded. These have included tariffs or import levies, import quotas, export subsidies, direct payments to farmers, and limitations on production. Tariffs and import quotas can be effective only if a nation normally imports some of its supply. Export subsidies result in higher prices to domestic consumers than to foreign purchasers; their use requires control over imports to prevent foreign supplies from entering the domestic market and bringing prices down. Direct payments to farmers have been used to maintain prices to consumers at reasonable levels, while assuring farmers a return above world-market levels. Limitations on production, intended to reduce supply and thus increase prices, have been used mainly in Brazil (for coffee) and in the United States (for major crops).

The U.S. approach. Since the enactment of the Agricultural Adjustment Act of 1933, the United States has had programs designed to limit the production of major farm crops through restrictions on acreage. Since that date it has also offered price supports for major crops such as wheat, feed grains, rice, tobacco, peanuts (groundnuts), and cotton, as well as for manufactured dairy products. It has not had price-support programs for perishable crops or for major livestock products except for a few years during and after World War II.

The Commodity Credit Corporation

The price-support method most widely used has been the nonrecourse loan made by the Commodity Credit Corporation (CCC): the farmer may repay the loan by delivering his produce at the support price or "redeem" it in cash if the market price is higher. The amount of particular crops offered for price-support loans has varied greatly from year to year, as have redemptions.

Most of the farm products given price supports were crops normally exported by the United States. Until the mid-1960s the price supports were above the export prices. Unless export subsidies were paid to make up the difference between domestic prices and the prices foreign buyers were willing to pay, exports became impossible. Export subsidies were accordingly paid on such farm commodities as cotton and grains. In the 1960s the support prices for the major export commodities, except tobacco, were established at levels near or slightly below world prices to permit market forces to manage the distribution of supplies between domestic and foreign markets. The lowering of support prices was accompanied by a substantial increase in the size of direct payments to farmers. By the end of the 1960s such payments had come to constitute a high percentage of the cash receipts from farm marketings: in cotton, 60 percent; in wheat, 40 percent; and in feed grains, 30 percent. These payments fell sharply in the late 1970s, largely as a result of increased demand.

In order to receive payments, farmers had to agree to limit the acreage devoted to specified crops. At the beginning of the 1970s the various programs had resulted in the diversion of approximately 20,000,000 hectares of land from the production of major farm crops. The number of acres diverted from cultivation fell sharply, however, in the late 1970s and early 1980s.

A major component of U.S. farm price policy since World War II has been the disposal of surplus produce abroad through the economic aid program. This began as an outgrowth of wartime Lend-Lease, and food exported from the United States made a major contribution to the postwar recovery of western Europe. The Agricultural Trade and Development Act of 1954 provided a base for continuing such activities, and gradually the emphasis shifted from western Europe to the developing countries. One of the important effects was to dispose of farm products that could not be sold either domestically or in regular commercial foreign trade. Without this the farm income and price objectives could not have been achieved except by more stringent output limitations, lower farm prices, and larger direct government payments.

The British approach. Efforts to control agricultural prices go far back in English history, although the early objectives were quite different from those of more recent times. The Corn Laws of the 15th century were designed to prevent prices from becoming too high; restrictions were imposed on the right to export corn (wheat) when the domestic price exceeded a specified level. In 1663 the laws were revised to prevent prices from falling too low, by including import duties when the home price did not exceed a specified level. The general trend, until the Corn Laws were finally abandoned in 1846, was increasingly toward ensuring higher prices for home producers through the payment of export bounties and by the restriction of imports until prices reached specified levels. After 1846 the British followed a free-trade policy for agricultural products but moved to the protection of agriculture and the establishment of minimum prices for certain farm products during the depression of the 1930s. Protection was expanded after World War II by legislation in 1947 and 1957 which sought to support farm prices primarily through deficiency payments to farmers, covering about 95 percent of total output. In most cases the domestic price was free to vary with changing demand and supply conditions; local products competed with imported supplies that were generally subject to relatively low tariffs. The farmer was reimbursed for the difference between his average realized price and a guaranteed price. The Agricultural Act of 1957, which gave the government the right to limit the amount of agricultural output on which deficiency payments were made, was designed to reduce the cost of the program and to encourage domestic production.

The British system of supporting farm prices, while allowing consumers the lowest possible food prices in the world market, was gradually abandoned during the late 1960s as the United Kingdom prepared for entry into the European Economic Community (EEC). When the United Kingdom entered the EEC in 1972, its agricultural prices began to rise to the much higher level prevailing within the EEC. The United Kingdom, moreover, imports more food and live animals from EEC countries than it exports, leading many British to question the value of membership in the EEC.

Policies of the EEC. The EEC has established a common agricultural policy (CAP) for the Common Market countries. The CAP, worked out for each major farm commodity, was originally designed to create free trade for that commodity within the community. Special subsidies by the individual countries, and other national farm programs, were to be eliminated to prevent competitive advantages. The first of the regulations implementing the CAP were enacted in 1962 and applied to grain (except rice), poultry, eggs, live hogs and whole hog carcasses, fruit and vegetables, and wine. Similar programs were developed later for beef, dairy products, sugar, rice, and fats and oils.

The regional approach

The most important features of the CAP mechanism are the target prices, the threshold prices, the support or intervention prices, the variable levies on imports to make up the difference between landed prices and threshold prices, and export subsidies or refunds equal to the difference between market prices in the EEC and in the importing country. For most CAP commodities the primary device for achieving target prices is the variable import levy. This levy, which fluctuates with the import cost of a commodity, keeps the domestic price at or near the target price if the commodity is imported. When EEC production of a commodity exceeds EEC consumption, the authorities may purchase the commodity for storage, pay to have it processed for another use (*e.g.*, wheat may be denatured and sold as a feed grain), or subsidize its export to countries outside the EEC. With these techniques the EEC has been able to maintain farm prices at levels substantially higher than those prevailing in the United States and Canada.

Throughout the 1960s the EEC did nothing to limit or control the production of agricultural products. When large stocks of butter and dry skim milk accumulated, and as the costs of maintaining dairy product prices and subsidizing wheat exports mounted, consideration was given to reducing production. A payments program to induce shifts from dairy to beef production was inaugurated, and there was talk of reducing the area cultivated for grain. Output limitation has been made difficult, however, by the significant differences in circumstances among the farmers in the various EEC countries.

The centralized approach

Soviet policies. The farm policies of the Soviet Union were established during the First Five-Year Plan (1928–32), when agriculture was collectivized. For all practical purposes, regular markets for farm products were abolished at that time, and each collective farm was required to deliver an assigned quota of produce to the state at very low prices. If a farm had anything left after meeting the obligatory quotas, it could sell the surplus to the state at higher prices or to the local free markets. Until 1958 the collective farms also had to make payments in kind to the machine tractor stations in return for work done.

After Stalin's death in 1953, farm prices were increased significantly; the average procurement prices for food products increased almost fourfold between 1950 and 1956. In 1958 the multiple-price system was abandoned, and the prices paid to collective farmers became almost seven times the average paid in 1950. The machine tractor stations were abolished in 1958 and the machinery transferred to individual farms. Another major revision of prices was made by the post-Khrushchev government in 1964–65. A two-price system replaced the single prices; prices for deliveries of grain up to the planned amount were about 10 percent higher than the previous single price, and deliveries above the plan level received a premium of 50 percent. Significant regional price differentials were established to cover the higher costs of production in some regions. Prices of livestock products had already been increased by about 35 percent in 1962, and in 1965 further increases of perhaps a third were made. Another important measure was a commitment that planned purchases were to be fixed at specified levels during the Eighth Five-Year Plan (1966–70), both in the aggregate and for individual farms. Prior to that time, if a farm had significantly increased its production or if other farms in the same region had failed to meet their deliveries, the delivery quota might be arbitrarily increased for the farm that happened to have had some output available for delivery.

Soviet price policy before 1953 was clearly designed to obtain farm products as cheaply as possible. The low prices were generally not passed on to consumers; a significant fraction of total governmental revenue was derived from high taxes on farm products. The changes made after 1953 were intended to provide farmers with an incentive to raise production and to make more efficient use of resources. Only a part of the increase in prices paid to farms was passed on to consumers; much of the increase was at the expense of government revenue.

In the late 20th century Soviet planning began to give greater emphasis to private plots; while constituting only about 1.5 percent of Soviet farmland, these plots produced about one-third of the nation's agricultural output other than grains. Restrictions on the crops private plots could produce were relaxed, and the importance of those plots was stressed. State farms and collectives, however, continued to receive the vast majority of capital and feed grains. Private plots, moreover, suffered from many of the problems that stunted the state farms and collectives, including the flight of young people from the countryside.

THE PROBLEM OF STANDARDS

None of the governments engaged in regulating farm prices and incomes has been able to apply a meaningful standard as to what a fair price or reasonable income is. The actual measures adopted, such as specific price supports or intervention prices, have been determined through the political process, with little reference to formal principles or standards.

Governmental attempts to regulate prices and incomes

In the United States the Agricultural Adjustment Act of 1933 stated that the goal should be to establish prices having the same purchasing power as those of the period 1910–14. By the end of World War II it had become clear that the "parity price" relationships of 1910–14 were no longer relevant to existing conditions. The Agricultural Act of 1948 retained the 1910–14 average as a parity for all farm products but stated that the parity for individual products was to be the average of prices over the most recent 10-year period. Since the application of this formula would have resulted in a significant reduction in the parity prices for some politically important farm products, particularly cotton and wheat, legislation that was passed in 1949 declared that the parity price for an individual commodity was to be determined by either the old or the new formula, whichever was higher. Not until 1955 were the "modernized parity" prices put into effect on a gradual basis. The Food and Agricultural Act of 1977 introduced cost of production as a standard for determining farm price supports. This standard, however, is far from absolute, since the cost of production varies from one region to another; moreover, many costs of production, such as rent, are influenced by the value of the crops produced.

There has been a similar lack of objective or measurable standards in other countries. In Great Britain, for example, the Agriculture Act of 1947 declared its intention to be that of

> promoting and maintaining a stable and efficient agricultural industry capable of producing such part of the nation's food and other agricultural produce as in the national interest it is desirable to produce in the United Kingdom, and of producing it at minimum prices consistently with proper remuneration and living conditions for farmers and workers in agriculture and an adequate return on capital invested in the industry.

Several other countries have legislation that aims, without specifying in practice what is meant, to obtain for farm people the same level of income as that of other groups in the economy or that states that farm people should share in the rise in real per capita incomes. Finland, Japan, France, Sweden, and Norway have such policy objectives. German legislation declares that agriculture should share in the progressive development of society and requires that the government each year prepare a report showing the extent to which the return to farm labour, or properly managed holdings under average conditions, is in line with that of wage earners in comparable nonagricultural occupations in rural areas.

The agricultural price objectives of the Treaty of Rome, which established the EEC, also lack practical significance:

> To ensure . . . a fair standard of living for the agricultural population, particularly by the increasing of the individual earnings of persons engaged in agriculture; to stabilize markets; to guarantee regular supplies; and to ensure reasonable prices in supplies to consumers.

ACCOMPLISHMENTS

The effects of price and income policies are difficult to assess. The policies have unquestionably worked to raise agricultural production in the countries where they have been applied, but their usefulness as a means of enhancing the economic well-being of farm people is debatable. The governments of the industrial countries have been able to raise the returns from agriculture above the levels that would have prevailed in the absence of such intervention. In addition to maintaining prices, they provide subsidies for agricultural inputs such as tractor fuel and chemical fertilizers; they also gave assistance in consolidating small farms into larger ones and in improving farm buildings. They do not, except for the United States, attempt to moderate the effects of these policies on production.

Factors affecting income level and economic well-being

The level of income and the economic well-being of farm people in general are determined by many factors, including not only the prices they receive for their output but also the rate at which the economy in general is growing, the ease with which people can move from farm to nonfarm jobs, the prices they must pay for their productive inputs, and their level of education. With respect to average income per person, as distinguished from total income, the prices received and paid are probably less important than the other factors mentioned. This becomes obvious when one compares farm incomes in the United States or the United Kingdom with those in Argentina or India; the differences in real income have to do mainly with the levels of economic development and not with farm prices or subsidies. Government efforts to increase farm prices are likely to be offset, in the long run, by an increase in the number of persons engaged in farming, and this tends to keep the returns to farm labour from rising much faster than they would in the absence of such policies.

There are two other reasons for believing that the income effects of higher farm prices or subsidies are relatively insignificant in the long run compared with other factors affecting incomes of farm workers. One is that an increase in farm prices induces farmers to use more fertilizer, machinery, fuel and oil, and other items. If a significant part of any increase in gross income is used for such things, the absolute increase in net farm income is much smaller than the increase in gross farm income. The second reason is that a given increase in government-supported farm prices generally occurs only once. After the increase in returns has been realized, the higher farm prices contribute nothing further to incomes. In contrast, general economic growth along with the continued reduction of the farm labour force has cumulative effects on the return to farm labour. If the returns to farm labour were to grow at an average annual rate of about 3 percent, for example, farm prices would have to increase at least 3 percent annually (assuming other prices did not change) to have the same effect on returns to farm resources.

COSTS

The costs of the agricultural price and income policies of industrial countries are substantial; they include not only direct governmental outlays but also the increased costs to consumers in those countries, as well as the losses to developing countries of potential export markets.

The high prices of farm products in the United States in the mid-1970s and the relaxation of interventionist policies by the EEC after 1974 substantially reduced the costs of farm programs in these two regions. With the decline of farm prices that began in 1976, costs to taxpayers and consumers again approached the levels of the early 1970s.

The organization of farming

OWNERSHIP

Except in nations with Communist governments, most farm land is privately owned. This does not mean, however, that the land is owned by those who farm it. In most countries a major aspiration of farm people has been to achieve the ownership of the land they work. After World War II, Japan and Taiwan underwent land reforms that were intended to broaden ownership; these are generally considered to have been highly successful. Similar reforms have been advocated in other countries.

The variety of forms of ownership and operation

On a cooperative farm the land is owned jointly by the members of the group who farm it. The cooperative generally also owns all the major means of production, and the members supply all or most of the labour. While there are examples of cooperative farms in many countries, they loom large only in Israel, where the kibbutzim control about a fifth of the total agricultural land.

In a collective farm, at least as organized in the former Soviet republics, the land was owned by the state but was permanently leased to the *kolkhoz* (collective farm). The *kolkhoz* owned its own equipment and livestock and was required to meet certain commitments to the state in the form of deliveries of farm products. In theory the members of the *kolkhoz* were to elect the officers of the farm and establish the procedures by which the net product was to be divided among the members for services performed. In practice, however, their autonomy was severely limited by the economic plans. In most cases these plans were incredibly detailed, specifying the crops to be grown, the times of plowing, planting, and harvesting, the quantities of fertilizer and manures to be used, and the kinds of livestock to be maintained.

On state farms the land and all other means of production are owned by the state. The workers are paid in wages, and management decisions are made by individuals directly responsible to the state.

KINDS OF FARM OPERATION

If a family farm is defined as one for which the farm operator and members of his family supply at least half of the labour, the majority of farms in the world are family farms. Family farming is carried on under a wide range of conditions, from the small farms of Asia to the highly mechanized farms of Canada, the United States, and the United Kingdom.

The family farm may be owned by the farmer or rented. The most rapidly expanding type of tenure in the United States is that in which the farmer owns part of the land and rents the remainder; almost one-third of all farmland in the United States consists of part-owner farms. This arrangement enables the farmer to increase the size of the farm through renting and to invest capital in machinery and livestock.

Family farms may be large in terms of total assets or sales. The relative importance of family farms among the largest farms in the United States has increased over the past few decades. One of the more striking changes in industrial countries has been the increased importance of nonfarm income received by farm families. In the United States, Canada, and Japan more than half of the total income of farm families comes from nonfarm sources, while in most western European countries at least a third of the income of farm families is earned outside of agriculture.

A system of tenant farming known as sharecropping developed in the South of the United States following the freeing of the slaves in the 19th century. It was essentially an adjustment of the plantation system created to permit the owners to maintain a large measure of control over farm operations. The sharecropper usually supplied only the labour, while the owner provided animal power, machinery, and most of the other inputs in the form of an advance. The sharecropper received what was left after he had paid back the owner—generally about half of what had been produced.

For various reasons, including the exodus of blacks from American agriculture, the introduction of farm machinery, and the reduction in the acreage of cotton, the number of sharecroppers in the South has diminished by well over 80 percent since 1935.

In the past several decades there has been a growth of large-scale farming run as a business enterprise. These "industrial farms" are of growing significance in world agriculture. There are farms covering extensive areas of land in Africa, South America, and Australia, but most of them do not rely heavily upon machinery or other purchased inputs. Farms in the United States are becoming larger as their numbers grow smaller. Such large farms tend to specialize in the production of vegetables, fruits, cotton, poultry and poultry products, and livestock.

COMPARATIVE STRENGTHS AND WEAKNESSES

If they were free to choose, most farm families would want to own the land they farm. Wherever collectivization of private farmers has been carried out, it has required the use of force or the threat of force. But if family farming is to be viable, it must function efficiently, which means that farmers must have access to adequate sources of credit; must be able to obtain fertilizers, machinery, and other equipment; and must be able to market their produce easily. Laws and institutions must be sufficiently flexible to permit the average size of farms to increase as economic growth occurs.

Collective farming did not fulfill the hopes of its early advocates. In the Soviet Union the collective farm was used by Stalin as a means of exploiting the rural population in order to finance the expansion of industrialization. In the post-Stalin era the incomes of collective farm members increased, and it was believed that many remaining difficulties could be eliminated if the farms were given greater freedom in running their affairs. Nothing in the concept of the collective farm required the imposition of delivery quotas, centralized control of farm investment, or a particular organization of farm labour. Another weakness of collective farms was the failure to provide adequate incentives for individual members. Because of the difficulties involved in rewarding members for their individual work on the common land, the household plots of the members all too often tended to flourish at the expense of the collective.

There is no ideal form of organization that fits all farming. Under some circumstances the ownership of land may absorb so much capital that other investments, such as machinery and livestock, are neglected. Land rental may be a better alternative for many families, especially those with limited capital. The Israeli kibbutz has made it possible for many people with little or no agricultural experience to learn farming techniques quickly and efficiently. The most important consideration is whether the other institutions—economic, political, and social—are adequate to provide farmers with a wide range of resources and alternatives. (D.G.J./Ed.)

THE BASIC UNIT OF OPERATION: THE FARM

Farm buildings

The basic unit of commercial agricultural operation, throughout history and worldwide, is the farm. Because farming systems differ widely, there are important variations in the nature and arrangements of farm facilities. The buildings on a farm generally consist of the farm family's house, the dwellings of any resident hired workers, and the various structures and facilities for farming operations. This section deals with farmhouses and service buildings that can be classified as follows: livestock barns and shelters; machinery- and supply-storage buildings; buildings and facilities for crop storage, including fodder; and special-purpose structures.

GENERAL LAYOUT

The location of the farmstead and the relative position of its different buildings are influenced by several factors, external and internal. Among the external factors, mainly natural, are soil conditions, climatic conditions, and access facilities to the main road and to the fields.

Internal factors depend on the type of business enterprise suitable to the farm. Among general principles that must be taken into account are the necessity of some partition between the farmhouse and service buildings, minimizing of transportation between buildings, the possibility of enlarging buildings, and security against fire. Four general layouts may be defined: large crop farms, large stock farms, farms in underdeveloped areas, and small to medium mixed farms.

Large crop farms. Independently owned farms of this type, mainly cash-grain farms, are numerous in North America. The layout is simple: there are generally two types of service buildings, one for storage and the other for machinery. Large farms specializing in fruit production have a shed for the conditioning and storing of products, the other main building being a machinery and supply shelter. Some large farms specializing in viticulture include buildings that are equipped with wine cellars.

Extensive and intensive stock farm types

Large stock farms. Two types of large stock farms, extensive and intensive, may be distinguished. The extensive type is exemplified by the cattle ranchers of the United States. At the extreme, there are no buildings, only equipment. In Australia and New Zealand, dairy cows are kept without housing. The only building houses the milking parlour and the milk room, in the centre of the pasture. In the western United States, the most important beef ranches have several thousand head, entirely free on the range. The only building is the elevator with the milling and mixing machinery. For the animals there are only troughs and fences. Among intensive stock farms are the big dairy units—with several hundred cows—in the United States, in western Europe (France, northern Italy), and in eastern Europe and the former Soviet republics. There are three major layouts: parallel buildings; monobloc buildings (in Hungary, for example); and circular layout, with the milking parlour in the centre (United States, northern Italy). The covered feedlots for fattening beef, in the U.S. Midwest and elsewhere, feed from several hundred to several thousand head of cattle and are generally built with a shelter for the animals and with tower or bunker silos. Large units for hog production frequently have many buildings, partly to reduce disease risks and partly to separate the various animals—for example, the suckling sows, in-pig sows, fattening pigs, and boars. Some systems, however, use only one or two types of buildings. Large poultry units, specialized either for egg or for broiler production, use large identical buildings, the number depending on the unit size.

Farms in underdeveloped areas. In the underdeveloped areas, two types of buildings are found: those of the latifundia, or large plantation-type farms, and those of the small-owner or tenant farms. In these, buildings are generally small and scattered, the construction of a single large building being too expensive.

Mixed farms. The small and medium farms, which characterize European agriculture and which exist in many other parts of the world, are managed on the traditional mixed-farming and animal-husbandry system. Consequently, this type of farm normally has several service buildings: one for machinery, one for hay and cattle, another for hogs, and still another for sheep. In mountain areas, however, there frequently is a single building, including the house. With the increase of the average size of farms in these areas, there is relative specialization, and the number of buildings in the newly built farms is decreasing.

BUILDING TYPES

These include homes (farmhouses), livestock barns and shelters, buildings for machinery and supplies, and crop-storage and special-purpose structures.

Farmhouses. The basic requirements for the farmer's family are about the same as those of the urban family, but certain features of the farmhouse depend on the farm-life pattern. Because the farmer generally comes directly from the fields or the service buildings, with soiled clothes and boots, it is necessary to provide a rear entrance with a washroom or lavatory and clothes-storage space. For the same reason, many farmers prefer a dining place close to the kitchen or included in it. The house must include an office and a large food-storage place with ample refrigeration, including a freezer in many countries, as most farm families are large. There are usually three or four bedrooms.

Satisfactory modernization of old farmhouses is difficult in some cases, but if the available floor space is sufficient and the main walls strong, renovation can give good results. The cost of a new house must be proportionate to the farmer's income; for this reason, farmhouses in underdeveloped regions have less floor space with a main room (kitchen and dining room), two or three bedrooms, a large washroom, and a storage place.

Livestock barns and shelters. These tend to become the most important elements of the farm layout. Two general types of animal shelters may be distinguished: the multipurpose type, a single-story building with clear-span roof construction, useful for feed storage and machinery, as well as for livestock; and the specific type, designed for a particular type of animal.

Types of cattle housing

There are two major cattle-housing methods, the stall barn (or stanchion barn) and the loose-housing system. In the stall barn each animal is tied up in a stall for resting, feeding, milking, and watering. The typical plan has two rows of stalls. In older buildings hay and straw are stored in an overhead loft, but in modern layouts adjacent buildings are generally used.

In cold and moderate climates the barns need insulated walls and ceilings, as well as ventilation systems, either natural or power-operated. In mild and hot areas the barns are open on one or two sides. The loose-housing system, developed in the United States after World War II, is now employed throughout the world. Basically, this system includes a wood- or metal-framed shelter, arranged in such a way that the animals can move freely inside and sometimes also between the shelter and an outside yard. Depending on the bedded areas, four types can be distinguished: loose housing on permanent litter—for example, straw, corncob, sawdust; loose housing in free stalls or cubicles; loose housing on slatted floors; and loose housing on sloped concrete.

In some countries, in old as well as in new buildings, dairy cows are housed in stall barns that include milk rooms. Milking takes place in stalls, and the milk is carried either in cans or directly by pipeline to a refrigerated tank in the milk room. Modern layouts with loose housing always include a milking parlour, either stationary or rotary. Two types of loose housing are used: loose housing on permanent litter and loose housing in free stalls, either under a clear-span roof or under a narrow lean-to roof. Beef-breeding cows often live on pastures, with only open-front sheds, during the calving period. In France and Scotland, however, they are kept in barns all winter. For fattening steers there are two major housing systems. The first of these is the American system, with very large groups of animals and a wide surface per animal. In the western United States the open feedlots include only fences, troughs, and alleys for feed distribution. In the Midwest Corn Belt a shelter is often included. The second, the European system, is characterized by very small groups (10 to 20 animals each) and a very small surface, generally covered. Any of the four loose-housing systems can be used.

For horses and ponies it is customary to use individual stalls, where the animal can move freely, even though this requires more space. Mules may be kept together in large pens. In mild climates sheep and goats live on pastures without any shelter. The facilities include fences, waterers, corrals, dipping vats, and lambing and shearing sheds. In moderate and cold climates the flock is wintered in sheds. The trend is toward clear-span buildings, with large alleys so that trailers can distribute feed into racks and troughs. Ewes are housed by groups (50 to 100 each), and special pens are kept for lambs. Feed racks and fence partitions are generally movable. For the dairy ewes there are special milking parlours. Goats are housed either in tie stalls, for small flocks under 50 head, with milking on the spot, or in pens, for larger flocks housed by groups, with milking in a special milking parlour. Pig housing varies for sows and fattening pigs. The sow lives with its litter for four to eight weeks according to the weaning age chosen. During this period there are two types of housing: movable, individual houses (generally of wood) located on or close to pastures and fixed in place, and central farrowing houses. A sow may farrow and live with its piglets in a single pen or farrow in a special stall, to avoid possibly crushing the piglets, or may farrow tied up by a chain or a harness. The pregnant sows live either free in groups of six to 12 or tied up or blocked up inside individual stalls. In cold climates the house is heated; in all modern practice infrared lamps or tubes are used to keep the piglets warm.

Fattening pigs, like fattening beef cattle, may be kept either in a simple feedlot, in large groups with a wide surface per head and a simple open shelter, a system widely used in the United States Corn Belt, or penned in a closed building, isolated and ventilated, each pen holding seven to 15 pigs. This is the most common system in Europe. Size of the pig units varies all the way from five sows or 20 pigs to large farms of up to 100,000 pigs. Poultry is the most industrialized type of animal production. Some of the breeding phases no longer take place in farms but in specialized plants; the farmer buys either chicks for broiler production or young layers for egg production. The typical modern broiler house holds from 10 to 100,000 birds, with automated feeding. Two types of facilities can be used. The broilers can be put on the ground on a deep litter of wood shavings, on wire mesh above a pit, or on a combination of these two floors. Alternatively, the broilers can be housed in metal cages, on three stories, each cage holding three to 10 animals. In this case, feeding and cleaning are mechanized and the density is higher. The typical laying house holds several thousand hens. The same facilities as for broilers are used, but use of the cage is more common for layers. There are several types of cages, some of which are mechanized to facilitate feeding, cleaning, and egg collecting. Each cage can hold one to five hens. The density can reach about 2 hens per square foot (23 hens per square metre). The main types are cages in two- or three-story batteries (California cages), which are not superposed but rise in tiers; and flat-deck cages, which allow maximum mechanization. The buildings are generally one story, fully enclosed; they have insulated structures with sophisticated ventilation systems. Turkeys and other fowl are housed like poultry but generally on the ground. Rabbit production involves housing by groups in cages, on one, two, or three stories.

Modern broiler houses

Buildings for machinery and supplies. This type of building is designed solely to afford protection from the weather, mainly rain. Machinery storage should have as much surface as possible between the interior posts, without being too deep, so that each machine can be taken out easily. The best solution is a clear-span shed, wood or metal-framed, 25 to 35 feet (eight to 10 metres wide), open on one side and 15 feet (4.5 metres) high under the gutter. At the end of the shed, one bay is reserved for repair and maintenance and another for tools. This part is equipped with sliding or overhead doors. The same shed, or another, can be used for storing the fertilizers, seeds, and pesticides.

Crop storage. Wheat, barley, shelled corn (maize), and other cereals can be stored in farm bins if the moisture is below a certain limit (from 10 to 15 percent). In some cases artificial drying is necessary before storage, though it is possible to store wet grain, especially shelled corn, in airtight silos for animal fodder. The most common methods of storage of dry grain are (1) in piles of five to 10 feet (1.5 to three metres) on a waterproof floor in a building with reinforced walls; (2) in square or round bins erected within a building, usually of timber, plywood, corrugated steel, or wire mesh lined with waterproof paper; and (3) in watertight bins, often of corrugated metal, with their own roofs, for outside erection. Ear corn is dried by natural ventilation through a crib of limited width, located in a building or outside. Loose or baled hay is stored and sometimes dried by ventilation with fresh or heated air, either under sheds or in special installations called hay towers. Silage is made to conserve moist fodders, such as corn, sorghum, and grass. There are two types of silos. The horizontal silo is a parallelepiped, either cut into the ground (trench silo) or built above ground (bunker silo). The floor is natural earth or concrete. The walls can be concrete, timber or plywood, or sheet steel. The capacity varies but can be large. The tower silo is an above ground cylinder, with 20- to 30-foot (six- to nine-metre) diameter and a 50- to 65-foot (15- to 20-metre) height.

Silos

Ordinary silos, which are only watertight, are of wood, concrete, masonry staves or blocks, or steel. Special airtight silos with steel walls and a fused-glass surface are used for storage of high dry-matter silage, called "haylage." Fruit and vegetable storage for family consumption is usually in caves or cellars. For crops to be marketed, conditioning and storage generally are handled by commercial enterprises, but some large, specialized farms have their own storage. The buildings are insulated, and temperature control is assured either by ventilation with outside air (*i.e.*, for potatoes and onions) or by refrigeration (*i.e.*, for apples).

Special-purpose structures. Many secondary farm structures, such as smokehouses and well houses, are a leftover of the past, but some are necessary in specialized farms. A typical example is the tobacco barn, built for static air circulation. (R.Ma./Ed.)

Farm management

Farm management normally consists of making and implementing the decisions involved in organizing and operating a farm for maximum production and profit. Farm management draws on agricultural economics for information on prices, markets, agricultural policy, and economic institutions such as leasing and credit. It also draws on plant and animal sciences for information on soils, seed, and fertilizer, on control of weeds, insects, and disease, and on rations and breeding; on agricultural engineering for information on farm buildings, machinery, irrigation, crop drying, drainage, and erosion control systems; and on psychology and sociology for information on human behaviour. In making his decisions, a farm manager thus integrates information from the biological, physical, and social sciences.

Because farms differ widely, the significant concern in farm management is the specific individual farm; the plan most satisfactory for one farm may be most unsatisfactory for another. Farm management problems range from those of the small, near-subsistence and family-operated farms to those of large-scale commercial farms where trained managers use the latest technological advances, and from farms administered by single proprietors to farms managed by the state.

The individual farm

In Southeast Asia the manager of the typical small farm with ample labour, limited capital, and only four to eight acres (1.6–3.2 hectares) of land, often fragmented and dispersed, faces an acute capital–land management problem. Use of early maturing crop varieties; efficient scheduling of the sequence of land preparation, planting, and harvesting; use of seedbeds and transplanting operations for intensive land use through multiple cropping; efficient use of irrigation and commercial fertilizer; and selection of chemicals to control insects, diseases, and weeds—all of these are possible measures for increasing production and income from each unit of land.

In western Europe the typical family farmer has less land than is economical with modern machinery, equipment, and levels of education and training, and so must select from the products of an emerging stream of technology the elements that promise improved crop and livestock yields at low cost; adjust his choice of products as relative prices and costs change; and acquire more land as farm labour is attracted by nonfarm employment opportunities and farm numbers decline.

On a typical 400-acre (160-hectare) corn-belt farm in the United States with a labour force equivalent to two full-time men, physical conditions and available technologies allow a wide range of options in farming systems. To reach a satisfactory income requires operating on an increasing scale of output and increasing specialization. Corn and soybean cash-crop farming systems have increased in number along with corn-hog-fattening farms and corn-beef-fattening farms. Thus, the choice of a farming system, the degree of specialization to be chosen, the size of operation, and the method of financing are top concerns of management.

For a typical crop-livestock farm in São Paulo's Paraíba Valley, Brazil, large-scale use of hired labour creates a substantial management problem. With 30 to 40 workers per establishment, procuring and managing the labour—keeping abreast of demand and supply conditions for hired labour, working out contractual arrangements (wage rates and other incentives), deciding how to combine labour with other inputs, and supervising the work force—are of critical importance.

A rancher with thousands of acres, whether in the pampas of Argentina, the plains of Australia, or the prairies of the United States, is concerned about the rate of increase of the herd through births and purchases and herd composition—cows, calves, yearlings, steers, heifers. Risks from drought, winter storms, and price changes can be high. Weather, prospective yields, and the price outlook are the constant concern of competent and alert farm managers.

The collective farm

On a collective farm in the Soviet Union with 30,000 acres (12,000 hectares) and 400 workers, major management decisions are made by party–state representatives; the collective-farm chairman responds largely to their directives, though the farm manager is being given greater autonomy. Major management concerns are determining optimal size of the collective, improving labour incentives, increasing crop and livestock yields, and reducing unit costs—with emphasis on levels of fertilizer, on pesticide and herbicide use, and on conservation of soil and water in crop production.

Thus, the character of the world's agriculture is shaped as millions of farmers manage the resources under their control in ways to obtain as much satisfaction as possible from their decisions and actions, which are made in a large variety of settings in regard to human, capital, and land resource combinations; technological possibilities; and social and political arrangements. Future agricultural progress depends on improving the quality of management and the environment in which farmers make decisions and on helping them adjust their decisions to the changing environment. In the low-income agricultures of the world in the 1980s, expanded research, improved input supplies and transport facilities, enlarged market opportunities, and an otherwise encouraging environment promise to open up a much wider area for managerial choice and decision making.

BASIC CONCERNS

Land, livestock, and labour. A good farm manager is familiar with the legal description of the farm property for which he is responsible, location relative to other property, roads, markets, and sources of supply, the details of the field arrangement and farmstead layout, the farm's capital position or relation of debts to assets, and the resources of the farm, such as the capabilities of its soils. Such facts enable the manager to analyze and evaluate his resources and plan their use. To calculate profit potential, the farm manager estimates the yield expected from each acre or hectare of land and from each head of livestock. He then applies money prices to these quantities.

Calculating farm size

The size of a farm business, an indication of its profit-making potential, is measured by the total number of acres or hectares in the farm, acres or hectares planted to cash crops, productive man-work units (the number of workdays of labour required under average efficiency to care for crops and livestock), livestock units kept, capital invested, and total cash receipts. While total acreage is often used to describe farm size, it is not a very satisfactory measure since it does not specify how much land is hilly, stony, swampy, or otherwise unproductive. Total cropped land, total receipts, invested capital, or productive work units are better measures. Though livestock are counted by the head for the sake of comparison, for management purposes one cow is roughly equal in value to two calves, five hogs, 10 young pigs, seven sheep, 14 lambs, or 100 laying hens.

While the amount of land in a farm is more or less fixed, many farmers buy or rent additional acreage to increase their volume of output as a means of reducing unit costs. If such acreage is available within a reasonable distance, then land can often be profitably exploited. Other ways of increasing volume include bringing unimproved pasture and woodland into the cropping plan and shifting either to more intensive methods of cultivation or to more valuable crops. Before making major changes, the farm manager attempts to assure himself that the new crops will grow well and will find a market in his area. Almost all the governments of the world today have departments or ministries of agriculture which have been established for the purpose of advancing agricultural welfare by spreading technological information. Often these agencies perform extensive experimentation with new crop varieties, new cultivation techniques, and improved breeds of livestock, thus reducing the burden of risk upon the individual farm manager contemplating such changes. Considerable experimentation and research are also carried out by private agricultural supply firms that hope to improve their competitive position in the marketplace by developing a valuable new product.

In some of the developing countries, traditional patterns of land tenure and laws of inheritance may result in one farmer holding many quite small plots at some distance from each other. To reduce the resulting labour inefficiency and low productivity and to spur development of large-scale agriculture, governments in these countries have frequently legislated to permit or compel consolidation of such holdings (see LAND REFORM AND TENURE).

Productive and non-productive farm labour

Some kinds of farm work are directly productive, some are indirectly productive, and some are not productive at all. Work such as plowing, planting, cultivating, harvesting, feeding, and milking is directly productive. Maintenance of fences, buildings, and machinery, though often necessary, is not directly productive. Such work as trimming shrubbery and mowing lawns, unless it adds to the market value of the farm, is not considered productive. Similarly, capital can be highly productive, as in the case of livestock; indirectly productive (*e.g.,* tractors, buildings, and supplies); or unproductive, as a large, showy barn or house. Land, too, can be highly productive, moderately so, or waste. Analysis of farm records has shown that farmers often overequip their property, thus using buildings and machinery to less than full capacity. Generally speaking, small farmers have been shown to have a higher proportion of their total investment in buildings than in machinery. In the developing countries, where relatively large quantities of human labour and relatively small amounts of capital are employed, a rather different problem exists. In these areas, farm managers need large numbers of people to work the fields during planting and harvest and far smaller numbers to perform routine cultivation tasks. In consequence, these countries face a problem of underemployment of agricultural labour during much of the year.

Financial management and large-scale operation. The financial tools a farmer can use to analyze, plan, and control his business include financial statements, profit and loss statements, and cash-flow statements. A financial statement tells the amount of money invested in farm assets, outstanding debts, the owner's equity in the business, and the degree to which the farm is liquid and solvent. Liquidity is the ability to meet financial obligations on time, whereas solvency is the ability to pay all debts if the business is forced to discontinue. A profit and loss statement shows sources and amounts of income and operating expenses. Comparison of profit and loss statements over a period of years tells which resources have been most profitable and whether there has been an advance or decline in net income. A cash-flow statement shows the sources and uses of funds at given periods during the year. Such a statement provides a useful check on the accuracy of the farm's other business records.

For the traditional farmer, land and labour (his own and that of his family) are the major resources. Under favour-

Conditions favourable to large-scale operations

able conditions, the farmer has changed his role from labourer to operator-manager; much larger farm units with high capital investments have resulted. Such conditions include the existence of a considerable body of applicable scientific knowledge, an opportunity for greater efficiency from large-scale operations, the existence of good markets and transportation, the opportunity to routinize and centrally direct farm work, and an absence of community antagonism to large-scale agriculture.

The trend to the substitution of capital for labour is especially noticeable in the United States, for example, where capital accounts for a steadily increasing proportion of farm inputs. In the United States in 1940, capital comprised 29 percent of farm inputs, labour 54 percent, and land 17 percent; by 1976 capital accounted for 62 percent of farm inputs, labour 16 percent, and land 22 percent. Capital typically replaces labour when large machines do the work of several men using smaller implements; when chemicals replace the scythe and hoe for weed control; when milking parlours, pipelines, and bulk tanks replace handmilking operations; when a mechanized installation replaces the fork and bushel basket in dairy, beef, or hog feeding; when automated sprinklers bring irrigation water to crops; when cisterns and lagoons handle animal waste; when combines and forced-air crop drying speed the harvesting of small grain; and in similar substitutions.

The technical knowledge that a modern large-scale farm manager must possess is frequently held to be far greater than that required of most businessmen with equal investment; the capital required to operate such a farm is beyond the reach of many. In consequence, financial-management techniques resembling those of industry are often employed. Capital is imported from the outside; production is scheduled to meet quantity, grade, and timing requirements; and labour is given specific tasks, as in a factory.

Recognizing the economic benefits of large-scale agriculture, many underdeveloped countries have attempted to create conditions for its existence. National governments, often with outside help, have financed large-scale development programs, involving irrigation or improvement of huge acreages by means of dams, drainage facilities, and canals, and these have revolutionized the lives of many traditional farm managers within the space of a few years. Improvements in crops and livestock, marketing techniques and organization, and transport and power have in some cases increased agricultural productivity and income several times over. Since capital and management have been in the hands of government, the traditional farm manager has, however, often lost some of his independence, and not all such programs have succeeded. Poor planning and management by government authorities and resistance from the farmers themselves have led to some expensive failures.

The commodity market

Reducing market risks. The marketplace for agricultural commodities is exceptionally risky for three important reasons. First, no single farm producer can place or withhold enough of a single item on the market to affect the market price; second, the quantity of a commodity taken off the market does not increase in proportion to price declines; third, the farm manager cannot respond to falling prices by quickly switching production from an unprofitable item to a profitable one. To reduce his risks and safeguard profits, the farm manager may specialize or diversify depending on conditions; he may also use the futures market (see below).

A specialized farm manager concentrates his effort on the production of one item such as wheat, cotton, milk, eggs, or fruit. By such specialization he can realize the benefits of large-scale production and can make the most money from an enterprise in which he is highly skilled. On the other hand, the specialist is vulnerable to sudden changes in the market, to plant and animal diseases, and to soil exhaustion resulting from cultivation of a single crop.

Diversification—the spreading of one's talents over more than one farming enterprise—may be accomplished horizontally or vertically. Horizontal diversification means the production of more than one item for sale. In vertical diversification, the farm manager handles raw products after harvest by processing, packaging, transporting, or even selling at retail. A poultry farmer who produces eggs and washes, candles, grades, packages, and markets them at retail is said to be vertically diversified. He has taken on some of the jobs that could have been performed elsewhere, and as a result he generally receives a better return for his efforts.

Programs of agricultural diversification have been carried out by some developing countries, with the government acting as a kind of national farm manager. Upon achieving independence, nations such as Ghana and Nigeria, in West Africa, found their economies highly dependent upon a single raw agricultural export (cocoa for Ghana; palm oil for Nigeria). Sharply falling prices for these commodities or epidemics of plant disease were seen to have disastrous effect on national prosperity. Erosion problems also caused concern. The governments responded by horizontally diversifying into other profitable crops and vertically diversifying in the establishment of industries to process these commodities or turn them into manufactured goods before export.

A capable farm manager may use the futures market to try to minimize his risks. In the futures market, the farm manager contracts with a buyer to deliver a given quantity of some commodity at a specified date in the future for an agreed price. The buyer is often a speculator who hopes that prices will rise, enabling him to sell the commodity or the contract at a profit. Futures markets enable the farm manager to establish in advance a price for a crop or earn payment for holding a crop in storage. Futures markets also permit some farmers to speculate on a price increase without storing a crop, establish in advance the price of livestock feed intended for later use, and establish an advance price for livestock.

SPECIAL CONCERNS OF SCALE

Farm management specifics vary all over the world; it is possible here to cite only some of the most typical practices in several leading agricultural countries.

Large-farm management. Research has shown that large farms produce more efficiently than small farms. In sugarcane production, for example, the most efficient farm may include many thousands of acres or hectares. Yet, a well-managed dairy farm might achieve greatest efficiency with two men and fewer than 100 cows. In the future, as technology advances, the farms that are managed most efficiently will probably be larger than the most efficient farms at present.

Large farms can reduce costs by claiming volume discounts on their purchases. They can negotiate prices on fertilizer, seed, crop chemicals, petroleum products, machinery, and repair services. Large operators also have an advantage in selling their products. Managers of large corn farms, for example, can contract directly with a large processor for an entire year's production of given quantity and quality for a specific date in the future, thus commanding a higher price. The middleman is eliminated, and production, handling, and processing can be prescheduled for greater efficiency. Large farms also have a smaller investment in machinery and buildings per crop acre.

United States. The increase in the capital requirements of United States farms has already been described above. These changes in American agriculture are, to a large degree, the result of a revolution in financial management. Up to about 1930, little outside capital was needed to finance farming operations. Today, capital investment has vastly increased; farmers obtain their production goods and services—land, machines, breeding stock, seed, fertilizer, and other necessities—in a variety of ways.

Meeting capital requirements

Renting land is one way. In contrast to earlier days when land ownership was considered the ideal, renting land is now a widely accepted management practice. Large acreages of corn land in the Corn Belt, wheat land in the Great Plains, and cotton land in California and Arizona are operated by renters. Renting land enables farmers to operate on a much larger scale than would be possible under ownership. Specialized rice growers in the Sacramento Valley of California, who own tractors, tillage tools, and harvesters, receive rice-acreage allotments from the federal government. Such growers own no land, renting it instead from owners who have no rice allotment. Growers prepare the ground, irrigate it with water supplied by the landowner, and contract for application of seed and fertilizer. When the crop is ripe, the growers harvest the rice with their own combines and haul it to a warehouse for drying and storage. In upland areas of the valley, other growers raise tomatoes under contract from a canner, renting their land from a general crop farmer.

Farmers who do not wish to tie up capital in high-priced farm machinery can contract for harvesting of such crops as wheat, corn, grain sorghum, and barley. An airplane operator may seed, fertilize, and apply weed spray for a rice grower. Vegetables, fruit, and nuts may be picked under contract by shipper-packers whose crews move from farm to farm. Similar operations in livestock include sheepshearing, dehorning, branding, and artificial insemination.

Renting machinery and livestock

Rental of machinery is another management device farmers use to obtain the services of equipment too expensive to be owned individually. Rental of livestock also is receiving attention. In the northeastern United States dairy farmers lease cows. The owner of the cows may be a contracting firm, a local bank, or an individual investor for whom the bank serves as agent. The scheme is useful both to older farmers who wish to retire but want to retain their interest in dairying and to young dairy farmers who want to expand but have limited capital.

Soviet Union. Following the Bolshevik Revolution of 1917, large landholdings were expropriated by the state and the land was distributed among the peasants. In 1928 collectivization of Soviet agriculture was initiated on a large scale; a three-part structure composed of state farms (*sovkhoz*), collective farms (*kolkhoz*), and private plots emerged. The state farms were owned, managed, and operated by the state. Workers on state farms were salaried employees of the state; farm managers were state appointees. During the 1960s and '70s state farms increased sharply in numbers. Much of the increase was the result of new state farms being established in the virgin land areas and the consolidation of smaller collective farms into state farms.

The collective farm leased land from the state and was worked by members of the collective under an elected committee that, as the management unit, had the responsibility of organizing land, labour, and capital in accordance with production requirements. For years, payment to collective members consisted of their share of the collective's produce or income from its sale. Each individual's share was determined by a workday unit that took into account the time spent performing a job and the level of skill required for the job. In the last few decades prior to the dissolution of the Soviet Union (in 1991), most collective farms had shifted to a monthly wage similar to that used by state farms.

Private plots up to two acres (0.8 hectare) in size and operated by individual workers occupied less than 3 percent of the planted area in the Soviet Union but produced nearly half the potatoes, 40 percent of the eggs, 20 percent of the meat, and 13 percent of the vegetables.

Though the Soviet farm manager's role did not include primary decision making, there was a trend from the 1960s toward more management autonomy in farm production. The Soviet government promoted greater efficiency in agriculture by increasing the level of inputs and by improving incentives to farm labourers. These measures included financial concessions to farmers and expanded use of fertilizers, pesticides, irrigation, and drainage. The Soviet farm manager performed additional functions that in other countries are carried out by government and welfare officials, such as providing roads, recreation, education, health care, and welfare to members of the collective.

Israel. A unique feature of the management of agriculture in Israel is its cooperative settlements, which evolved as a result of the needs encountered by immigrants who were new both to their surroundings and to farming as a profession.

The two basic types of cooperative settlement are the moshav and kibbutz. A moshav is a village containing up to 150 farm family units and supported by a strong multipurpose cooperative organization. Each family is an economic and social unit, living in its own house and managing and working its own fields. Although each farm family is independent, its social and economic security is ensured by the cooperative structure of the village, whose organization markets the produce, purchases the farm and household equipment, and provides the farmer with credit and other services.

A kibbutz, numbering from 60 to 2,000 members, is a true collective based on common ownership of resources and on pooling of labour and income; it functions as a single democratic unit. Under the supervision of a manager, each member performs an assigned task but receives no salary or wages, because all the members' needs are provided by the kibbutz.

Israel's agriculture is highly organized into farm societies. One society, the Farmer's Federation, has a membership of 7,000 citrus growers. There are plantation development companies and associations of wine, fruit, milk, and cotton producers.

Production for export

Australia. A significant characteristic of farm management in Australia is the emphasis on production for export markets. Since the production of fine wool is the most important rural industry, grazing of sheep is a leading enterprise. Production of wheat, meat, dairy products, and fruit for export also figures large in the nation's agricultural economy. Australian export production is highly organized through statutory marketing authorities. Ten such authorities supervise the marketing of wheat, dairy products, meat, eggs, canned fruits, dried fruits, apples and pears, wine, honey, and wool.

Getting started in almost any farming venture in Australia requires substantial amounts of capital.

Management of small and middle-sized farms. *Canada.* Canadian agriculture consists largely of family farms, managed and operated by the owners. Less than one farm in 100 has hired management. A Canadian farm may vary in size from a factory-type broiler chicken plant of an acre or two (up to one hectare) to a cattle ranch that includes several townships. On a mechanized grain farm a farmer may operate 1,000 acres (400 hectares) or more with very little hired help. While most farmers in Canada own the farms they operate, there is a growing tendency to rent additional land. Current management trends also include increased use of commercial fertilizer and chemicals for pest control.

Farm management practices vary widely. Some farmers who rent land pay cash rent. In other cases the landlord takes a share of the crop or a share of the income from the sale of livestock or milk. On farms where most of the income is derived from the sale of grain, it is common for the tenant to give the landlord one-third of all grain produced. The landlord supplies the land, pays the taxes and fire insurance on the buildings, and provides materials for maintaining buildings and fences. Integration, the management of two or more stages of production and marketing, is spreading, with the trend most noticeable with sugar beets and canning crops.

United Kingdom. British farmers are well known for their efficient management and use of mechanical aids. Milking machines are employed on all but the smallest farms; electricity is widespread; grain combines are common; and there is one tractor for about every 35 acres (14 hectares) of arable land. British farmers also use great quantities of commercial fertilizer per acre, the cost of which is subsidized by the government. The government also subsidizes the cost of lime, eradication of tuberculosis, and construction of silos and other capital equipment and pays part of the cost of voluntary consolidation of small farms into more efficient commercial units.

Several agricultural commodities are subject to the authority of government marketing boards: some buy produce, others control producer–buyer contracts, and still others maintain broad control over marketing conditions. Cereals, potatoes, eggs, sugar beets, and wool are the principal products governed by marketing bodies.

Denmark. In Denmark successful farmer cooperatives play a major management role, extending credit and controlling production, marketing, import, export, purchasing, and sales. Through these cooperatives, Danish farmers enjoy the benefits of large-scale production and distribution despite the small size of individual farms. About 90 percent of Denmark's output of pork and milk and about 50 percent of egg output is marketed cooperatively. The number of farms in Denmark has been declining in recent years, but those remaining are becoming larger. Average size in the late 1970s was 54 acres (22 hectares). The family farm predominates.

Farm management in developing countries. *India.* Farm management practices in India range from the modern and sophisticated to some that have been in use for centuries. Illiteracy, inadequate water, unreliable power supplies, poor transportation and communications, making the timely acquistition of supplies and marketing of produce difficult—all hamper the development of modern farm management practices. For example, many farmers are unable to read the directions on a sack of fertilizer, to write an application for a production loan, or to calculate their profit and loss. Where progress has been made in introducing improved farm management techniques, visual and oral methods of instruction and training are being used successfully. Training techniques include on-farm demonstrations, farmer exchange programs, tours, short courses, literacy classes, exhibits, and audio–visual vans.

Republic of Zaire. Shifting cultivation is the typical method of farming in the Republic of Zaire. The native farmer clears two or three acres (about one hectare) in the forest or savanna, crops it until the fertility of the land declines, then moves on to another area. Fertilizer, insecticides, and fungicides are not generally available.

A land-settlement plan, called the paysannat system, in which strips of cultivated land were alternated with bush and grassland, was introduced in the 1930s to increase production. This system, however, has disintegrated since independence due to the lack of management personnel and government extension services and disruption of marketing channels. Often side by side with traditional farms are large modern plantations owned, managed, and operated by individual Europeans and corporations. Plantation crop yields are two to 10 times those of indigenous farms, probably pointing the direction of future development.

(M.E.Bl.)

PRINCIPLES AND PRACTICES OF PLANT CULTIVATION

Plant breeding

In its modern form, plant breeding involves the application of genetic principles to produce plants that are more useful to man. This is accomplished by selecting plants that man finds desirable, economically or aesthetically, by controlling the mating of the selected individuals and by selecting desirable individuals among the progeny. These processes, repeated over many generations, can change the hereditary makeup and value of a plant population far beyond the natural limits of previously existing populations. This article emphasizes the application of genetic principles to the improvement of plants; the biological factors underlying plant breeding are dealt with in the article GENETICS AND HEREDITY.

Historical background

Plant breeding is an ancient activity of man, dating back to the very beginnings of agriculture. Probably soon after the earliest domestications of cereal grains, man began to recognize degrees of excellence among the plants in his fields and saved seed from the best plants for planting new crops. Such tentative selective methods were the forerunners of early plant-breeding procedures.

The results of early plant-breeding procedures were conspicuous. Most present-day varieties are so modified from their wild progenitors they are unable to survive in nature. Indeed, in some cases, the cultivated forms are so strikingly different from existing wild relatives that it is even difficult to identify their ancestors. These remarkable transformations were accomplished by early plant breeders in a very short time from an evolutionary point of view, and the rate of change was probably greater than for any other evolutionary event.

Scientific plant breeding dates back hardly more than 50 years. The role of pollination and fertilization in the process of reproduction was not widely appreciated even 100 years ago, and it was not until the early part of the 20th century that the laws of genetic inheritance were recognized and a beginning was made toward applying them to the improvement of plants. One of the major facts that has emerged during the short history of scientific breeding is that an enormous wealth of genetic variability exists in the plants of the world and that only a start has been made in tapping its potential.

GOALS

The plant breeder usually has in mind an ideal plant that combines a maximum number of desirable characteristics. These characteristics may include resistance to diseases and insects; tolerance to heat and frost; appropriate size, shape, and time to maturity; and many other general and specific traits that contribute to improved adaptation to the environment, ease in growing and handling, greater yield, and better quality. The breeder of fancy show plants must also consider aesthetic appeal. Thus the breeder can rarely focus attention on any one characteristic but must take into account the manifold traits that make the plant more useful in fulfilling the purpose for which it is grown.

Increase of yield. One of the aims of virtually every breeding project is to increase yield. This can often be brought about by selecting obvious morphological variants. One example is the selection of dwarf, early maturing varieties of rice. These dwarf varieties are sturdy and give a greater yield of grain. Furthermore, their early maturity frees the land quickly, often allowing an additional planting of rice or other crop the same year.

Another way of increasing yield is to develop varieties resistant to diseases and insects. In many cases the development of resistant varieties has been the only practical method of pest control. Perhaps the most important feature of resistant varieties is the stabilizing effect they have on production and hence on steady food supplies. Varieties tolerant to drought, heat, or cold provide the same benefit.

Modifications of range and constitution. Another common goal of plant breeding is to extend the area of production of a crop species. A good example is the modification of grain sorghum since its introduction to the United States about 100 years ago. Of tropical origin, grain sorghum was originally confined to the southern Plains area and the Southwest, but earlier maturing varieties were developed until grain sorghum is now an important crop as far north as North Dakota.

Varieties suited to mechanized farming

Development of crop varieties suitable for mechanized agriculture has become a major goal of plant breeding in recent years. Uniformity of plant characters is very important in mechanized agriculture because field operations are much easier when the individuals of a variety are similar in time of germination, growth rate, size of fruit, and so on. Uniformity in maturity is, of course, essential when crops such as tomatoes and peas are harvested mechanically.

The nutritional quality of plants can be greatly improved by breeding. For example, it is possible to breed varieties of corn (maize) much higher in lysine than previously existing varieties. Breeding high-lysine maize varieties for those areas of the world where maize is the major source of this nutritionally essential amino acid has become a major goal in plant breeding.

In breeding ornamentals, attention is paid to such factors as longer blooming periods, improved keeping qualities of flowers, general thriftiness, and other features that contribute to usefulness and aesthetic appeal. Novelty itself is often a virtue in ornamentals, and the spectacular, even the bizarre, is often sought.

EVALUATION OF PLANTS

The appraisal of the value of plants so that the breeder can decide which individuals should be discarded and which allowed to produce the next generation is a much more difficult task with some traits than with others.

Qualitative characters. The easiest characters, or traits, to deal with are those involving discontinuous, or qualitative, differences that are governed by one or a few major genes. Many such inherited differences exist, and they frequently have profound effects on plant value and utilization. Examples are starchy versus sugary kernels (characteristic of field and sweet corn, respectively) and determinant versus indeterminant habit of growth in green beans (determinant varieties are adapted to mechanical harvesting). Such differences can be seen easily and evaluated quickly, and the expression of the traits remains the same regardless of the environment in which the plant grows. Traits of this type are termed highly heritable.

Quantitative characters. In other cases, however, plant traits grade gradually from one extreme to another in a continuous series, and classification into discrete classes is not possible. Such variability is termed quantitative. Many traits of economic importance are of this type; *e.g.*, height, cold and drought tolerance, time to maturity, and, in particular, yield. These traits are governed by many genes, each having a small effect. Although the distinction between the two types of traits is not absolute, it is nevertheless convenient to designate qualitative characters as those involving discrete differences and quantitative characters as those involving a graded series.

Controlling quantitative characters

Quantitative characters are much more difficult for the breeder to control, for three main reasons: (1) the sheer numbers of the genes involved make hereditary change slow and difficult to assess; (2) the variations of the traits involved are generally detectable only through measurement and exacting statistical analyses; and (3) most of the variations are due to the environment rather than to genetic endowment; for example, the heritability of certain traits is less than 5 percent, meaning that 5 percent of the observed variation is caused by genes and 95 percent is caused by environmental influences.

It follows that carefully designed experiments are required to distinguish plants that are superior because they carry desirable genes from those that are superior because they happen to grow in a favourable site.

METHODS OF PLANT BREEDING

Mating systems. Plant mating systems devolve about the type of pollination, or transferal of pollen from flower to flower (see the article REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Pollination*). A flower is self-pollinated (a "selfer") if pollen is transferred to it from any flower of the same plant and cross-pollinated (an "outcrosser" or "outbreeder") if the pollen comes from a flower on a different plant. About half of the more important cultivated plants are naturally cross-pollinated, and their reproductive systems include various devices that encourage cross-pollination; *e.g.*, protandry (pollen shed before the ovules are mature, as in the carrot and walnut), dioecy (stamens and pistils borne on different plants, as in the date palm, asparagus, and hops), and genetically determined self-incompatibility (inability of pollen to grow on the stigma of the same plant, as in white clover, cabbage, and many other species).

Other plant species, including a high proportion of the most important cultivated plants such as wheat, barley, rice, peas, beans, and tomatoes, are predominantly self-pollinating. There are relatively few reproductive mechanisms that promote self-pollination; the most positive of which is failure of the flowers to open (cleistogamy), as in certain violets. In barley, wheat, and lettuce the pollen is shed before or just as the flowers open; and in the tomato pollination follows opening of the flower, but the stamens form a cone around the stigma. In such species there is always a risk of unwanted cross-pollination.

Control of breeding

In controlled breeding procedures it is imperative that pollen from the desired male parent, and no other pollen, reaches the stigma of the female parent. When stamens and pistils occur in the same flower, the anthers must be removed from flowers selected as females before pollen is shed. This is usually done with forceps or scissors. Protection must also be provided from "foreign" pollen. The most common method is to cover the flower with a plastic or paper bag. When the stigma of the female parent becomes receptive, pollen from the desired male parent is transferred to it, often by breaking an anther over the stigma, and the protective bag is replaced. The production of certain hybrids is, therefore, tedious and expensive because it often requires a series of delicate, exacting, and properly timed hand operations. When male and female parts occur in separate flowers, as in corn (maize), controlled breeding is easier.

A cross-pollinated plant, which has two parents, each of which is likely to differ in many genes, produces a diverse population of plants hybrid (heterozygous) for many traits. A self-pollinated plant, which has only one parent, produces a more uniform population of plants pure breeding (homozygous) for many traits. Thus, in contrast to outbreeders, self-breeders are likely to be highly homozygous and hence true breeding for a specified trait.

Breeding self-pollinated species. The breeding methods that have proved successful with self-pollinated species are (1) mass selection; (2) pure-line selection; (3) hybridization, with the segregating generations handled by the pedigree method, the bulk method, or by the backcross method; and (4) development of hybrid varieties.

Mass selection. In mass selection, seeds are collected from (usually a few dozen to a few hundred) desirable appearing individuals in a population, and the next generation is sown from the stock of mixed seed. This procedure, sometimes referred to as phenotypic selection, is based on how each individual looks. Mass selection has been used widely to improve old "land" varieties, varieties that have been passed down from one generation of farmers to the next over long periods.

An alternative approach that has no doubt been practiced for thousands of years is simply to eliminate undesirable types by destroying them in the field. The results are similar whether superior plants are saved or inferior plants are eliminated: seeds of the better plants become the planting stock for the next season.

A modern refinement of mass selection is to harvest the best plants separately and to grow and compare their progenies. The poorer progenies are destroyed and the seeds of the remainder are harvested. It should be noted that selection is now based not solely on the appearance of the parent plants but also on the appearance and performance of their progeny. Progeny selection is usually more effective than phenotypic selection when dealing with quantitative characters of low heritability. It should be noted, however, that progeny testing requires an extra generation; hence gain per cycle of selection must be double that of simple phenotypic selection to achieve the same rate of gain per unit time.

Mass selection, with or without progeny test, is perhaps the simplest and least expensive of plant-breeding procedures. It finds wide use in the breeding of certain forage species, which are not important enough economically to justify more detailed attention.

Pure-line selection. Pure-line selection generally involves three more or less distinct steps: (1) numerous superior appearing plants are selected from a genetically variable population; (2) progenies of the individual plant selections are grown and evaluated by simple observation, frequently over a period of several years; and (3) when selection can no longer be made on the basis of observation alone, extensive trials are undertaken, involving careful measurements to determine whether the remaining selections are superior in yielding ability and other aspects of performance. Any progeny superior to an existing variety is then released as a new "pure-line" variety. Much of the success of this method during the early 1900s depended on the existence of genetically variable land varieties that were waiting to be exploited. They provided a rich source of superior pure-line varieties, some of which are still represented among commercial varieties. In recent years the pure-line method as outlined above has decreased in importance in the breeding of major cultivated species;

Exploitation of sports

however, the method is still widely used with the less important species that have not yet been heavily selected. A variation of the pure-line selection method that dates back centuries is the selection of single-chance variants, mutations or "sports" in the original variety. A very large number of varieties that differ from the original strain in characteristics such as colour, lack of thorns or barbs, dwarfness, and disease resistance have originated in this fashion.

Hybridization. During the 20th century planned hybridization between carefully selected parents has become dominant in the breeding of self-pollinated species. The object of hybridization is to combine desirable genes found in two or more different varieties and to produce pure-breeding progeny superior in many respects to the parental types.

Genes, however, are always in the company of other genes in a collection called a genotype. The plant breeder's problem is largely one of efficiently managing the enormous numbers of genotypes that occur in the generations following hybridization. As an example of the power of hybridization in creating variability, a cross between hypothetical wheat varieties differing by only 21 genes is capable of producing more than 10,000,000,000 different genotypes in the second generation. At spacings normally used by farmers, more than 50,000,000 acres would be required to grow a population large enough to permit every genotype to occur in its expected frequency. While the great majority of these second generation genotypes are hybrid (heterozygous) for one or more traits, it is statistically possible that 2,097,152 different pure-breeding (homozygous) genotypes can occur, each potentially a new pure-line variety. These numbers illustrate the importance of efficient techniques in managing hybrid populations, for which purpose the pedigree procedure is most widely used.

Pedigree breeding starts with the crossing of two genotypes, each of which have one or more desirable characters lacked by the other. If the two original parents do not provide all of the desired characters, a third parent can be included by crossing it to one of the hybrid progeny of the first generation (F_1). In the pedigree method superior types are selected in successive generations, and a record is maintained of parent–progeny relationships.

The F_2 generation (progeny of the crossing of two F_1 individuals) affords the first opportunity for selection in pedigree programs. In this generation the emphasis is on the elimination of individuals carrying undesirable major genes. In the succeeding generations the hybrid condition gives way to pure breeding as a result of natural self-pollination, and families derived from different F_2 plants begin to display their unique character. Usually one or two superior plants are selected within each superior family in these generations. By the F_5 generation the pure-breeding condition (homozygosity) is extensive, and emphasis shifts almost entirely to selection between families. The pedigree record is useful in making these eliminations. At this stage each selected family is usually harvested in mass to obtain the larger amounts of seed needed to evaluate families for quantitative characters. This evaluation is usually carried out in plots grown under conditions that simulate commercial planting practice as closely as possible. When the number of families has been reduced to manageable proportions by visual selection, usually by the F_7 or F_8 generation, precise evaluation for performance and quality begins. The final evaluation of promising strains involves (1) observation, usually in a number of years and locations, to detect weaknesses that may not have appeared previously; (2) precise yield testing; and (3) quality testing. Many plant breeders test for five years at five representative locations before releasing a new variety for commercial production.

Bulk-population breeding

The bulk-population method of breeding differs from the pedigree method primarily in the handling of generations following hybridization. The F_2 generation is sown at normal commercial planting rates in a large plot. At maturity the crop is harvested in mass, and the seeds are used to establish the next generation in a similar plot. No record of ancestry is kept. During the period of bulk propagation natural selection tends to eliminate plants having poor survival value. Two types of artificial selection also are often applied: (1) destruction of plants that carry undesirable major genes and (2) mass techniques such as harvesting when only part of the seeds are mature to select for early maturing plants or the use of screens to select for increased seed size. Single plant selections are then made and evaluated in the same way as in the pedigree method of breeding. The chief advantage of the bulk population method is that it allows the breeder to handle very large numbers of individuals inexpensively.

Often an outstanding variety can be improved by transferring to it some specific desirable character that it lacks. This can be accomplished by first crossing a plant of the superior variety to a plant of the donor variety, which carries the trait in question, and then mating the progeny back to a plant having the genotype of the superior parent. This process is called backcrossing. After five or six backcrosses the progeny will be hybrid for the character being transferred but like the superior parent for all other genes. Selfing the last backcross generation, coupled with selection, will give some progeny pure breeding for the genes being transferred. The advantages of the backcross method are its rapidity, the small number of plants required, and the predictability of the outcome. A serious disadvantage is that the procedure diminishes the occurrence of chance combinations of genes, which sometimes leads to striking improvements in performance.

Hybrid vigour

Hybrid varieties. The development of hybrid varieties differs from hybridization in that no attempt is made to produce a pure-breeding population; only the F_1 hybrid plants are sought. The F_1 hybrid of crosses between different genotypes is often much more vigorous than its parents. This hybrid vigour, or heterosis, can be manifested in many ways, including increased rate of growth, greater uniformity, earlier flowering, and increased yield, the last being of greatest importance in agriculture.

By far the greatest development of hybrid varieties has been in corn (maize), primarily because its male flowers (tassels) and female flowers (incipient ears) are separate and easy to handle, thus proving economical for the production of hybrid seed. The production of hand-produced F_1 hybrid seed of other plants, including ornamental flowers, has been economical only because greenhouse growers and home gardeners have been willing to pay high prices for hybrid seed.

Recently, however, a built-in cellular system of pollination control has made hybrid varieties possible in a wide range of plants, including many that are self-pollinating, such as sorghums. This system, called cytoplasmic male sterility, or cytosterility (Figure 1), prevents normal maturation or function of the male sex organs (stamens) and results in defective pollen or none at all. It obviates the need for removing the stamens either by hand or by machine. Cytosterility depends on the interaction between male sterile genes ($R + r$) and factors found in the cytoplasm of the female sex cell. The genes are derived from each parent in the normal Mendelian fashion, but the cytoplasm (and its factors) is provided by the egg only; therefore, the inheritance of cytosterility is determined by the female parent. All plants with fertile cytoplasm produce viable pollen, as do plants with sterile cytoplasm but at least one R gene; plants with sterile cytoplasm and two r genes are male sterile (produce defective pollen).

The production of F_1 hybrid seed between two strains is accomplished by interplanting a sterile version of one strain (say A) in an isolated field with a fertile version of another strain (B). Since strain A produces no viable pollen, it will be pollinated by strain B, and all seeds produced on strain A plants must therefore be F_1 hybrids between the strains. The F_1 hybrid seeds are then planted to produce the commercial crop. Much of the breeder's work in this process is in developing the pure-breeding sterile and fertile strains to begin the hybrid seed production.

Breeding cross-pollinated species. The most important methods of breeding cross-pollinated species are (1) mass selection; (2) development of hybrid varieties; and (3) development of synthetic varieties. Since cross-pollinated species are naturally hybrid (heterozygous) for many traits and lose vigour as they become purebred (homozygous),

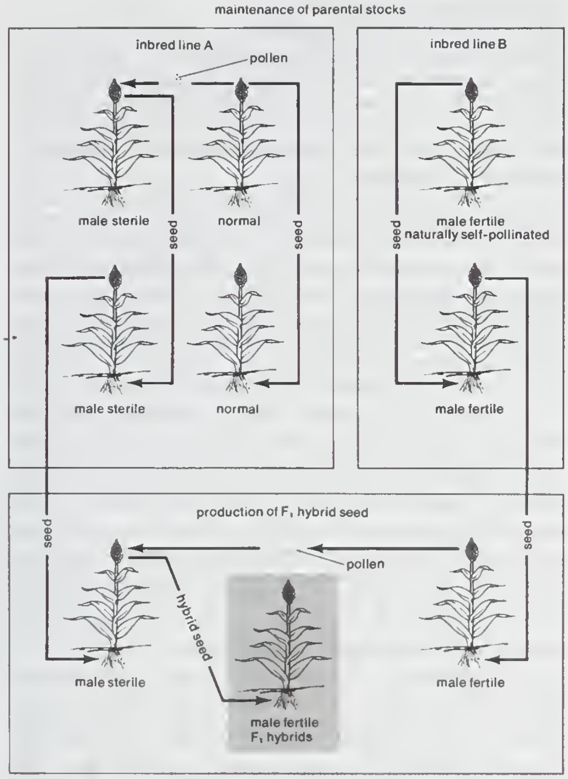


Figure 1: Production of hybrid seed in sorghum using cytoplasmic male sterility.

a goal of each of these breeding methods is to preserve or restore heterozygosity.

Mass selection. Mass selection in cross-pollinated species takes the same form as in self-pollinated species; *i.e.*, a large number of superior appearing plants are selected and harvested in bulk and the seed used to produce the next generation. Mass selection has proved to be very effective in improving qualitative characters, and, applied over many generations, it is also capable of improving quantitative characters, including yield, despite the low heritability of such characters. Mass selection has long been a major method of breeding cross-pollinated species, especially in the economically less important species.

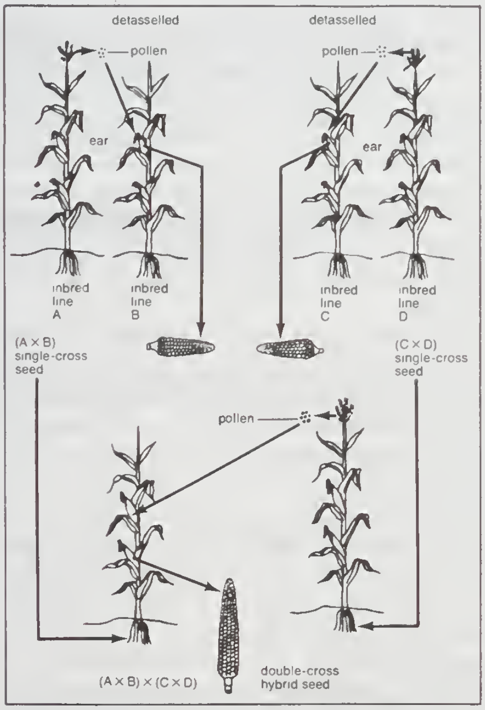


Figure 2: The crossing of inbred lines of corn (maize) to produce single crosses and the crossing of single crosses to produce double-cross hybrid seed.

Production of hybrid corn

Hybrid varieties. The outstanding example of the exploitation of hybrid vigour through the use of F_1 hybrid varieties has been with corn (maize). The production of a hybrid corn variety (Figure 2) involves three steps: (1) the selection of superior plants; (2) selfing for several generations to produce a series of inbred lines, which although different from each other are each pure-breeding and highly uniform; and (3) crossing selected inbred lines. During the inbreeding process the vigour of the lines decreases drastically, usually to less than half that of field-pollinated varieties. Vigour is restored, however, when any two unrelated inbred lines are crossed, and in some cases the F_1 hybrids between inbred lines are much superior to open-pollinated varieties. An important consequence of the homozygosity of the inbred lines is that the hybrid between any two inbreds will always be the same. Once the inbreds that give the best hybrids have been identified, any desired amount of hybrid seed can be produced.

Pollination in corn (maize) is by wind, which blows pollen from the tassels to the styles (silks) that protrude from the tops of the ears. Thus controlled cross-pollination on a field scale can be accomplished economically by interplanting two or three rows of the seed parent inbred with one row of the pollinator inbred and detasselling the former before it sheds pollen. In practice most hybrid corn is produced from "double crosses," in which four inbred lines are first crossed in pairs (A × B and C × D) and then the two F_1 hybrids are crossed again (A × B) × (C × D). The double-cross procedure has the advantage that the commercial F_1 seed is produced on the highly productive single cross A × B rather than on a poor-yielding inbred, thus reducing seed costs. In recent years cytoplasmic male sterility, described earlier, has been used to eliminate detasselling of the seed parent, thus providing further economies in producing hybrid seed.

D F Jones, *Genetics*, no 9 (1924)

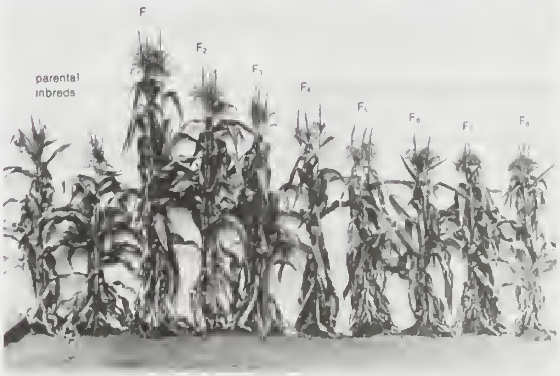


Figure 3: *Heterosis followed by decline in vigour in corn (maize).*
Representative plants of the two parental inbreds are at left. To their right are vigorous F_1 hybrids and the F_2 through F_8 selfed generations. The decline in vigour from F_1 through F_8 is similar to that observed when plants from open-pollinated varieties are self-pollinated.

Much of the hybrid vigour exhibited by F_1 hybrid varieties is lost in the next generation. Consequently, seed from hybrid varieties is not used for planting stock but the farmer purchases new seed each year from seed companies.

Perhaps no other development in the biological sciences has had greater impact on increasing the quantity of food supplies available to the world's population than has the development of hybrid corn (maize). Hybrid varieties in other crops, made possible through the use of male sterility, have also been dramatically successful and it seems likely that use of hybrid varieties will continue to expand in the future.

Synthetic varieties. A synthetic variety is developed by intercrossing a number of genotypes of known superior combining ability—*i.e.*, genotypes that are known to give superior hybrid performance when crossed in all combinations. (By contrast, a variety developed by mass selection is made up of genotypes bulked together without having undergone preliminary testing to determine their performance in hybrid combination.) Synthetic varieties are known for their hybrid vigour and for their ability to produce usable seed for succeeding seasons. Because of these advantages, synthetic varieties have become increasingly favoured in the growing of many species, such as the forage crops, in which expense prohibits the development or use of hybrid varieties.

Distribution and maintenance of new varieties. The benefits of superior new varieties obviously cannot be realized until sufficient seed has been produced to permit commercial production. Although the primary function of the plant breeder is to develop new varieties, he usually also carries out an initial small-scale seed increase. Seed thus produced is called breeders seed. The next stage is the multiplication of breeders seed to produce foundation seed. Production of foundation seed is usually carried out by seed associations or institutes, whose work is regulated by government agencies. The third step is the production of certified seed, the progeny of foundation seed, produced on a large scale by specialized seed growers for general sale to farmers and gardeners. Certified seed must be produced and handled in such a way as to meet the standards set by the certifying agency (usually a seed association). Seed associations are also usually responsible for maintaining the purity of new varieties once they have been released for commercial production.

Breeders, foundation, and certified seed

The distribution of new varieties developed by commercial plant-breeding companies is often through seed associations, but many reputable companies market their products without following the official certification process. In some countries, particularly in Europe, new varieties can be patented for periods up to 15 years or more, during which time the breeder has an exclusive right to reproduce and sell the variety. (R.W.A.)

Irrigation and drainage

Irrigation is the artificial application of water to land. Drainage is the artificial removal of excess water from land. Though either practice may be, and both often are, used for nonagricultural purposes to improve the environment, this section is limited to their major application, to agriculture. Some land requires irrigation or drainage before it is possible to use it for any agricultural production; other land profits from either practice to increase agricultural production. Some land, of course, does not need either practice.

Irrigation and drainage improvements are not necessarily mutually exclusive. Often both may be required together to assure sustained, high-level production of crops.

Riwkin

Figure 4: One of the modern concrete-lined canals of the Israel National Water Carrier that carries water from Lake Tiberias in the north to the Negev, the arid lands of the south.

MODERN IRRIGATION-SYSTEM PLANNING AND CONSTRUCTION

Water supply. The first consideration in planning an irrigation project is developing a water supply. Water supplies may be classified as surface or subsurface. Though both surface and subsurface water come from precipitation such as rain or snow, it is far more difficult to determine the origin of subsurface water.

In planning a surface water supply, extensive studies must be made of the flow in the stream or river that will be used. If the streamflow has been measured regularly over a long period, including times of drought and flood, the studies are greatly simplified. From streamflow data, determinations can be made of the minimum, maximum, average daily, and average monthly flows; the size of dams, spillways, and downstream channel; and the seasonal and carry-over storage needed. If adequate streamflow data are not available, the streamflow may be estimated from rain and snow data, or from flow data from nearby streams that have similar climatic and physiographic conditions.

Studying streamflow

The quality, as well as the quantity, of surface water is a factor. The two most important considerations are the amount of silt carried and the kind and amount of salts dissolved in the water. If the silt content is high, sediment will be deposited in the reservoir, increasing maintenance costs and decreasing useful life periods. If the salt concentration is high, it may damage crops or accumulate in the soil and eventually render it unproductive.

Subsurface sources of water must be as carefully investigated as surface sources. In general, less is known about subsurface supplies of water than about surface supplies, so, therefore, subsurface supplies are harder to investigate. Engineers planning a project need to know the extent of the basic geological source of water (the aquifer), as well as the amount the water level is lowered by pumping and the rate of recharge of the aquifer. Often the only way for the engineer to obtain these data reliably is to drill test wells and make onsite measurements. Usually, a project is planned so as not to use more subsurface water than is recharged. Otherwise, the water is said to be "mined," meaning that as a natural resource it is being used up.

Two sources of water not often thought of by the layman are dew and sewage. In certain parts of the world, Israel and part of Australia, for example, where atmospheric conditions are right, sufficient dew may be trapped at night to provide water for irrigation. Elsewhere the supply of waste water from some industries and municipalities is sufficient to irrigate relatively small acreages. Recently, due to greater emphasis on purer water in streams, there has been increased interest in this latter practice.

In some countries (Egypt for example) sewage is a valuable source of water. In others, such as the United States, irrigation is looked upon as a means of disposing of sewer water as a final step in the waste-treatment process. Unless the water contains unusual chemical salts, such as sodium, it is generally of satisfactory quality for agricultural irrigation. Where the practice is used primarily as a means of disposal, large areas are involved and the choice of crop is critical. Usually only grass or trees can withstand the year-round applications.

Determining water rights

Before a water supply can be assured, the right to it must be determined. Countries and states have widely varying laws and customs that determine ownership of water. If the development of a water supply is for a single purpose, then the determination of ownership may be relatively simple; but if the development is multipurpose, as most modern developments are, ownership may be difficult to determine, and agreements must be worked out among countries, states, municipalities, and private owners.

The area that can be irrigated by a water supply depends on the weather, the type of crop grown, and the soil. Numerous methods have been developed to evaluate these factors and predict average annual volume of rainfall needed. Some representative annual amounts of rainfall needed for cropland in the western United States are 12 to 30 inches (305 to 760 millimetres) for grain and 24 to 60 inches (610 to 1,525 millimetres) for forage. In the Near East, cotton needs about 36 inches (915 millimetres), whereas rice may require two to three times that amount.

In humid regions of the United States, where irrigation supplements rainfall, grain crops may require six to nine inches (150 to 230 millimetres) of water. In addition to satisfying the needs of the crop, allowances must be made for water lost directly to evaporation and during transport to the fields.

Transport systems. The type of transport system used for an irrigation project is often determined by the source of the water supply. If a surface water supply is used, a large canal or pipeline system is usually required to carry the water to the farms because the reservoir is likely to be distant from the point of use. If subsurface water drawn from wells is used, a much smaller transport system is needed, though canals or pipelines may be used. The transport system will depend as far as possible on gravity flow, supplemented if necessary by pumping. From the mains, water flows into branches, or laterals, and finally to distributors that serve groups of farms. Many auxiliary structures are required, including weirs (flow-diversion dams), sluices, and other types of dams. Canals are normally lined with concrete to prevent seepage losses, control weed growth, eliminate erosion hazards, and reduce maintenance. The most common type of concrete canal construction is by slip forming. In this type of construction, the canal is excavated to the exact cross section desired and the concrete placed on the earth sides and bottom.

Water pipelines

Pipelines may be constructed of many types of material. The larger lines are usually concrete whereas laterals may be concrete, cement–asbestos, rigid plastic, aluminum, or steel. Although pipelines are more costly than open conduits, they do not require land after construction, suffer little evaporation loss, and are not troubled by algae growth.

Water application. After water reaches the farm it may be applied by surface, subsurface, or sprinkler-irrigation methods. Surface irrigation is normally used only where the land has been graded so that uniform slopes exist (see Figure 5). Land grading is not necessary for other methods. Each method includes several variations, only the more common of which are considered here.

Surface irrigation systems are usually classed as either flood or furrow systems. In the flood system, water is applied at the edge of a field and allowed to move over the entire surface to the opposite side of the field. Grain and forage crops are quite often irrigated by flood techniques. The furrow system is used for row crops such as corn (maize), cotton, sugar beets, and potatoes. Furrows are plowed between crop rows and the water is run in the furrows. In either type of surface irrigation systems, wastewater ditches at the lower edge of the fields permit excess water to be removed for use elsewhere and to prevent waterlogging.

By courtesy of U.S. Soil Conservation Service

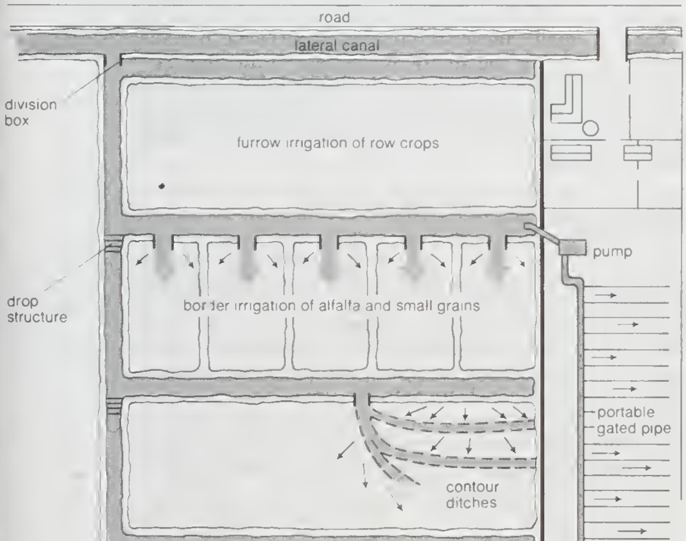

road
lateral canal
division box
furrow irrigation of row crops
drop structure
border irrigation of alfalfa and small grains
pump
portable gated pipe
contour ditches
waste ditch
border ditches
basins for flooding level ground

Figure 5: Surface methods of applying irrigation water to field crops.

Subirrigation is a less common method. An impermeable layer must be located below, but near, the root zone of the crop so that water is trapped in the root zone. If this condition exists, water is applied to the soil through tile drains or ditches.

In recent years sprinklers have been used increasingly to irrigate agricultural land. Little or no preparation is needed, application rates can be controlled, and the system may be used for frost protection and the application of chemicals. Sprinklers range from those that apply water in the form of a mist to those that apply an inch or more per hour.

Evaporation and seepage control. Various techniques have been tried to reduce losses of irrigation water. Two major sources of loss, particularly from surface supplies and surface systems, are evaporation and seepage from reservoirs and canals. Many studies have been made of techniques to suppress evaporation. One of the more promising appears to be application of a special alcohol film on the surface, which retards evaporation by about 30 percent and does not reduce the quality of the water. The primary problem in its use is that it is fragile; a strong wind can blow it apart and expose the water to evaporation.

Seepage has largely been controlled by lining main and distribution channels with impervious material, typically concrete. Other materials used are asphalt and plastic film, though plastic tends to deteriorate if it is exposed to sunlight.

Typical systems. The typical surface irrigation system utilizes a publicly developed water supply—*e.g.*, a river-basin reservoir. The public project also constructs the main canals to take water from the reservoir to the agricultural land. In general the canals flow by gravity, but lift stations are often required. Supply and field canals are used to bring the water to the individual field, where it is applied to the land either by furrow or by flooding method.

Sprinkler irrigation

Until recently most sprinkler-irrigation systems depended on privately developed water supplies, but many modern sprinkler systems have been able to draw on public water supplies. In either case, a pump is required to pump water from a large (1,000 gallons, or 3,785 litres, per minute and larger) well or a supply canal. The water goes into the system main and thence to a sprinkler unit. Many automatic or semiautomatic moving sprinkler systems travel over the field applying water. Two common units are the so-called centre pivot and the travelling sprinkler. The centre-pivot unit is anchored at the centre of the field; a long lateral (arm) with sprinklers mounted on it sweeps the field in a circle. The system has the disadvantage of missing the corners of a square field. A travelling sprinkler is mounted on a trailer and propelled across the field in a lane that has been left unplanted. The unit drags a flexible hose connected to the main supply line. When it reaches the end of the lane, it is automatically shut off and can be moved to the next lane. Despite some shortcomings, all sprinkler systems are effective in applying a controlled amount of water at a high level of efficiency with a minimum of labour.

MODERN DRAINAGE-SYSTEM PLANNING AND CONSTRUCTION

Planning a system. The planning and design of drainage systems is not an exact science. Although there have been many advances in soil and crop science, techniques have not been developed for combining the basic principles involved into precise designs. One of the primary reasons for difficulty in applying known theory is the capricious variability of natural soil in contrast to the idealized soils required to develop a theory.

The type of drainage system designed depends on many factors, but the most important is the type of soil, which determines whether water will move through rapidly enough to use subsurface drainage. Soils that have a high percentage of sand- and silt-size particles and a low percentage of clay-size particles usually will transmit water rapidly enough to make subsurface drainage feasible. Soils that are high in clay-size particles usually cannot be drained by subsurface improvements. It is essential to consider soil properties to a depth of five to six feet (1.5 to 1.8 metres) because the layer in the soil that transmits water the slowest controls the design, and subsurface improvements may be installed to these depths.

The topography or slope of the land is also important. In many cases, land in need of drainage is so flat that a contour map showing elevations 12 inches (30 centimetres) or six inches (15 centimetres) apart is used to identify trouble spots and possible outlets for drainage water. Often an outlet can be developed only by collective community action. The rainfall patterns, the crops to be grown, and the normal height of the water table also are considered. If heavy rainfall is not probable during critical stages of crop growth, less extensive drainage improvements may suffice. The capacity of the system is governed in part by the growth pattern of the crop, its planting date, critical stages of growth, tolerance of excess water, harvest date, and value.

In some areas the normal water level in the soil is high, in others low; this variable is always investigated before a drainage system is planned.

Types of drainage systems. Drainage systems may be divided into two categories, surface and subsurface. Each has several components with similar functions but different names. At the lower, or disposal, end of either system is an outlet. In order of decreasing size, the components of a surface system are the main collection ditch, field ditch, and field drain; and for a subsurface system, main, submain, and lateral conduits from the submain. The outlet is the point of disposal of water from the system; the main carries water to the outlet; the submain or field ditch collects water from a number of smaller units and carries it to the main; and the lateral or field drain, the smallest unit of the system, removes the water from the soil.

Drainage system outlet

The outlet for a drainage system may be a natural stream or river or a large constructed ditch. A constructed ditch usually is trapezoidal in section with side banks flat enough to be stable. Grass may be grown on the banks, which are kept clear of trees and brush that would interfere with the flow of water.

A surface drainage system removes water from the surface of the soil and to approximately the bottom of the field ditches. A surface system is the only means for drainage improvement on soils that transmit water slowly. Individual surface drains also are used to supplement subsurface systems by removing water from ponded areas.

The field drains of a surface system may be arranged in many patterns. Probably the two most widely used are parallel drains and random drains. Parallel drains are channels running parallel to one another at a uniform spacing of a few to several hundred feet apart, depending on the soil and the slope of the land. Random drains are channels that run to any low areas in the field. The parallel system provides uniform drainage, whereas the random system drains only the low areas connected by channels. In either case the channels are shallow with flat sides and may be farmed like the rest of the field. Crops are usually planted perpendicular to the channels so that the water flows between the rows to the channels.

Some land grading of the fields where surface drains are installed is usually essential for satisfactory functioning. Land grading is the shaping of the field so that the land slopes toward the drainage channels. The slope may be uniform over the entire field or it may vary from part to part. Before the advent of the digital computer, the calculations necessary for planning land grading were time-consuming, a factor that restricted the alternatives available for final design. Today, computers rapidly explore many possibilities before a final land grading design is selected.

In a subsurface drainage system, often called a tile system, all parts except the outlet are located below the surface of the ground. It provides better drainage than a surface system because it removes water from the soil to the depth of the drain, providing plants a greater mass of soil for root development, permitting the soil to warm up faster in the spring, and maintaining a better balance of bacterial action, the air in the soil, and other factors needed for maximum crop growth.

The smallest component of the subsurface system, the lateral, primarily removes water from the soil. The laterals may be arranged in either a uniform or a random pattern. The choice is governed by the crop grown and its value, the characteristics of the soil, and the precipitation pattern.

Depth of spacing of laterals

The primary decision required for a system with uniform laterals is their depth and spacing. In general, the deeper the laterals can be emplaced, the farther apart they can be spaced for an equivalent degree of drainage. Theoretical studies have shown that laterals can be spaced 24 feet (7.3 metres) apart for each foot of depth. Laterals usually are spaced from 80 to 300 feet (24 to 91 metres) apart and three to five feet (0.9 to 1.5 metres) deep.

Subsurface drainage systems are as important in many irrigated areas as they are in humid areas. A drainage system is needed on irrigated lands to control the water table and ensure that water will be able to move through a soil, thus keeping salts from accumulating in the root zone and making the soil unproductive.

Construction and maintenance. Most subsurface drains are constructed by excavating a trench, installing a tile, and backfilling the trench. Work is in progress in the United States and in Europe to develop machines that will install drain tubes without excavating the trench. Control of the machines to assure proper slope of the drain has been a major problem, but recent development in excavation technology, including the use of laser beams for grade control, have helped to solve it. Traditionally, clay or concrete tile has been the principal material used, but many types of perforated plastic tubes are now employed. An advantage is the reduction in weight of the material handled.

With proper maintenance, drainage systems give relatively long life. Selected herbicides are applied to keep woody growth and water weeds out of the channels. Grates are usually installed over outlets to prevent rodents and burrowing animals from building nests.

Surface drainage systems need almost yearly maintenance to assure the slope and cross section of the channels and the slope of the graded areas because the slopes are so flat that small changes in the ground surface can make marked changes in the ability of a system to function.

Subsurface systems need periodic inspection but usually require little servicing. The outlet of the system and infrequent structural failure of the material are the usual points for service.

LAND RECLAMATION THROUGH IRRIGATION AND DRAINAGE

The need for increased food and fibre production in the 1980s and '90s requires the continued development of new agricultural lands. Development of such land is rarely possible without irrigation or drainage systems or both. Easily recognized improvements are the large-scale river-basin projects designed for flood control, irrigation, and power generation. Such projects are in various stages of design or construction in many countries of the world—for example, the People's Republic of China, India, Egypt, Iran, Australia, and the United States. In almost all cases drainage of the irrigated lands is considered a companion requirement. If possible, the drainage improvements are subsurface.

Reclaiming salty soils

A combination of drainage and irrigation is being used to reclaim large areas of land that have been abandoned because of salt accumulation. In this case subsurface drainage systems must be installed so that high water tables are lowered and pure water flushed through the soil, dissolving the salts and carrying them away in the drainage water. Large areas in the United States, India, and the Middle East are potentially available for reclamation by this technique.

The people of The Netherlands have reclaimed land from the sea by the use of drainage. Since the IJsselmeer (formerly Zuiderzee) barrier dam was closed in 1932, converting this large body of water into a freshwater lake, the Dutch have been continually enclosing and reclaiming smaller bodies (polders). After dikes are built around a polder, the area is drained by pumping out the water. Drainage channels and, in many places, subsurface drains are installed so that the root zone of crops can be drained. After this, cropping is started as the last step in the reclamation process.

The development of land-clearing machinery and surface-drainage techniques has made it possible to clear and drain tropical lands for agricultural production. The first step is the removal of trees, brush, and other tropical growth. Outlet ditches are constructed, followed by drains. In some cases subsurface drains are possible, but more often the soils and rainfall conditions combine to make this improvement impractical. Surface drains are installed on a uniform pattern and the land is smoothed or graded. Drainage systems on newly reclaimed tropical land require special attention while the soils are stabilizing, and some reconstruction is often needed after the soil stabilization is complete.

IRRIGATION AND DRAINAGE THROUGHOUT THE WORLD

The Food and Agriculture Organization of the United Nations (FAO) keeps the most complete statistics on irrigated lands; it estimates that in the entire world some 520,000,000 acres (211,700,000 hectares) are irrigated. FAO data, supplied by each country, indicate that the largest areas under irrigation are located in such countries as the People's Republic of China, India, Pakistan, and the United States. More than 130 countries report some acreage under irrigation. The largest area reported was estimated as 113,700,000 acres (46,000,000 hectares) in the People's Republic of China. Asia, excluding the former Soviet republics, irrigates close to 65 percent of the total area of the world that is irrigated; most of this is the large surface-irrigated, rice-producing areas of the People's Republic of China, India, Pakistan, and Southeast Asia. The United States has approximately 10 percent of the world's irrigated areas. Europe has roughly 7 percent, South America and Africa each about 4 percent, and Central America about 3 percent. Australia and New Zealand together have 1 percent or less. Sprinkler irrigation is employed throughout the world, but the largest acreage to make use of the sprinkler method is the approximately 9,900,000 acres (4,006,500 hectares) in the United States.

Statistics on drainage improvements are sparser than statistics on irrigation. It may safely be said that drainage in one form or another is practiced in almost every country of the world. It is now universally accepted that drainage is needed as much on irrigated as on nonirrigated land. Coantries such as India that have large-scale river-basin developments planned with irrigation also have companion drainage systems planned so that the land will not be damaged by salt accumulation.

U.S. drainage census

Some indication of the world picture may be gained from the drainage census in the United States of 1959, which showed that about 92,000,000 acres (37,200,000 hectares) were drained through organized projects, about 10 percent of the land in agriculture. A rule of thumb states that there is at least one acre of privately drained land for each acre in an organized project, indicating about 185,000,000 acres (75,000,000 hectares) of agricultural land drained in the United States at that time. In the late 1970s about 5 percent of the agricultural land in the United States was drained.

It is almost certain that the land area of the world improved by irrigation and drainage will continue to increase because these practices are two of the most elemental means of reclaiming and improving agricultural lands.

(B.A.J.)

Soil preparation

Mechanical processing of soil so that it is in the proper physical condition for planting is usually referred to as tilling; adding nutrients and trace elements is called fertilizing. Both processes are important in agricultural operations.

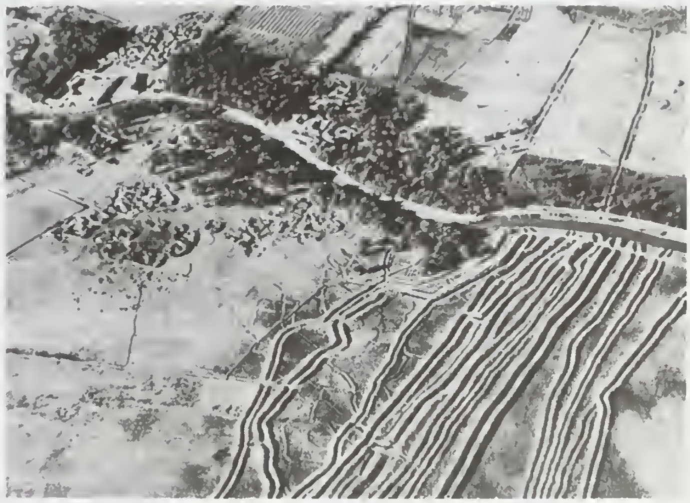


Figure 6: Inundation canals running from the Euphrates River, Iraq, that fill when the water level of the river rises.
Popperfoto

TILLING

Tillage is the manipulation of the soil into a desired condition by mechanical means; tools are employed to achieve some desired effect (such as pulverization, cutting, or movement). Soil is tilled to change its structure, to kill weeds, and to manage crop residues. Soil-structure modification is often necessary to facilitate the intake, storage, and transmission of water and to provide a good environment for seeds and roots. Elimination of weeds is important, because they compete for water, nutrients, and light. Crop residues on the surface must be managed in order to provide conditions suitable for seeding and cultivating a crop.

Desirable soil structure

Generally speaking, if the size of the soil aggregates or particles is satisfactory, preparation of the seedbed will consist only of removing weeds and the management of residues. Unfortunately, the practices associated with planting, cultivating, and harvesting usually cause destruction of soil structure. This leaves preparation of the seedbed as the best opportunity to create desirable structure, in which large and stable pores extend from the soil surface to the water table or drains, ensuring rapid infiltration and drainage of excess or free water and promoting aeration of the subsoil. When these large pores are interspersed with small ones, the soil will retain and store moisture also.

Seedbed-preparation procedures depend on soil texture and the desired change in size of aggregates. In soils of coarse texture, tillage will increase aggregate size, provided it is done when only the small pores are just filled with water; tillage at other than this ideal moisture will make for smaller aggregates. By contrast, fine-textured soils form clods; these require breakage into smaller units by weathering or by machines. If too wet or too dry, the power requirements for shattering dry clods or cutting wet ones are prohibitive when using tillage alone. Thus, the farmer usually attempts tillage of such soils only after a slow rain has moistened the clods and made them friable.

Some soils require deepening of the root zone to permit increased rate of water intake and improved storage. Unfavourable aeration in zones of poor drainage also limits root development and inhibits use of water in the subsoil.

Tillage, particularly conventional plowing, may create a hardpan, or plow sole; that is, a compacted layer just below the zone disturbed by tillage. Such layers are more prevalent with increasing levels of mechanization; they reduce crop yields and must be shattered, allowing water to be stored in and below the shattered zone for later crops.

Primary tillage equipment. Equipment used to break and loosen soil for a depth of six to 36 inches (15 to 90 centimetres) may be called primary tillage equipment. It includes moldboard, disk, rotary, chisel, and subsoil plows.

The moldboard plow

The moldboard plow is adapted to the breaking of many soil types. It is well suited for turning under and covering crop residues. There are hundreds of different designs, each intended to function best in performing certain tasks in specified soils. The part that breaks the soil is called the bottom or base; it is composed of the share, the landside, and the moldboard.

When a bottom turns the soil, it cuts a trench, or furrow, throwing to one side a ribbon of soil that is called the furrow slice. When plowing is started in the middle of a strip of land, a furrow is plowed across the field; on the return trip, a furrow slice is lapped over the first slice. This leaves a slightly higher ridge than the second, third, and other slices. The ridge is called a back furrow. When two strips of land are finished, the last furrows cut leave a trench about twice the width of one bottom, called a dead furrow. When land is broken by continuous lapping of furrows, it is called flat broken. If land is broken in alternate back furrows and dead furrows, it is said to be bedded or listed.

Different soils require different-shaped moldboards in order to give the same degree of pulverization of the soil. Thus, moldboards are divided into several different classes, including stubble, general-purpose, general-purpose for clay and stiff-sod soil, slat, blackland, and chilled general-purpose (Figure 7). The blackland bottom is used, for example, in areas in which the soil does not scour easily; that is, where the soil does not leave the surface of the emerging plow clean and polished.

By courtesy of John Deere

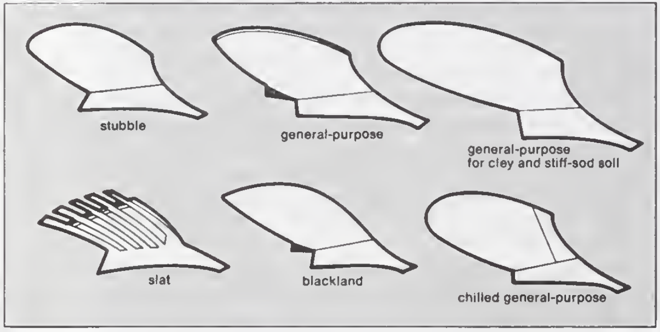


Figure 7: Plow bottom types.

The share is the cutting edge of the moldboard plow. Its configuration is related to soil type, particularly in the down suction, or concavity, of its lower surface. Generally, three degrees of down suction are recognized: regular for light soil, deep for ordinary dry soil, and double-deep for clay and gravelly soils. In addition, the share has horizontal suction, which is the amount its point is bent out of line with the landside. Down suction causes the plow to penetrate to proper depth when pulled forward, while horizontal suction causes the plow to create the desired width of furrow.

Moldboard-plow bottom sizes refer to width between share wing and the landside. Tractor-plow sizes generally range from 10 to 18 inches (25 to 45 centimetres), although larger, special-purpose types exist.

On modern mechanized farms, plow bottoms are connected to tractors either as trailing implements or integrally. One or more bottoms may be so attached (Figure 8). They are found paired right and left occasionally (two-way), with the advantage of throwing the furrow slice in a constant direction as the turns are made. A variation is the middlebreaker, or lister, which is a bottom equipped with both right- and left-handed moldboards.

The disk plow

The disk plow employs round, concave disks of hardened steel, sharpened and sometimes serrated on the edge, with diameters ranging from 20 to 38 inches (50 to 95 centimetres). It reduces friction by making a rolling bottom in place of a sliding one. Its draft is about the same as that of the moldboard plow. The disk plow works to advantage in situations where the moldboard will not, as in sticky nonscouring soils; in fields with a plow sole; in dry, hard ground; in peat soils; and for deep plowing. The disk-plow bottom is usually equipped with a scraper that aids in pulverizing the furrow slice. Disk plows are either trailed or mounted integrally on a tractor.

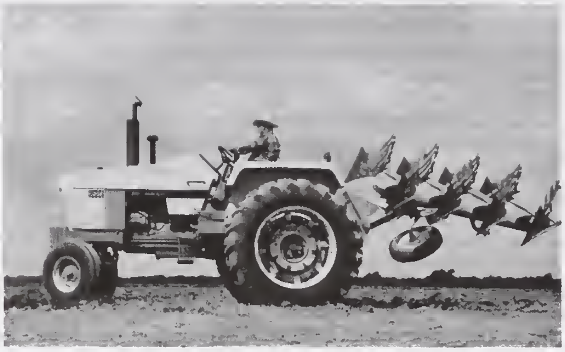


Figure 8: Tractor equipped with integrally mounted four-bottom two-way slat plow.
By courtesy of Robert E Stewart

The rotary plow's essential feature is a set of knives or tines rotated on a shaft by a power source. The knives chop the soil up and throw it against a hood that covers the knife set. These machines can create good seedbeds, but their high cost and extra power requirement have limited general adoption, except for the small garden tractor.

The chisel plow is equipped with narrow, double-ended shovels, or chisel points, mounted on long shanks. These points rip through the soil and stir it but do not invert and pulverize as well as the moldboard and disk plows. The chisel plow is often used to loosen hard, dry soils prior to using regular plows; it is also useful for shattering plow sole.

Subsoil plows are similar in principle but are much larger, since they are used to penetrate soil to depths of 20 to 36 inches (50 to 90 centimetres). Tractors of 60 to 85 horsepower are required to pull a single subsoil point through a hard soil at a depth of 36 inches. These plows are sometimes equipped with a torpedo-shaped attachment for making subsurface drainage channels.

Harrows and pulverizers

Secondary tillage. Secondary tillage, to improve the seedbed by increased soil pulverization, to conserve moisture through destruction of weeds, and to cut up crop residues, is accomplished by use of various types of harrows, rollers, or pulverizers, and tools for mulching and fallowing. Used for stirring the soil at comparatively shallow depths, secondary-tillage equipment is generally employed after the deeper primary-tillage operations; some primary tillage tools, however, are usable for secondary tillage. There are five principal types of harrows: the disk, the spike-tooth, the spring-tooth, the rotary cross-harrow, and the soil surgeon. Rollers, or pulverizers, with V-shaped wheels make a firm and continuous seedbed while crushing clods. These tools often are combined with each other.

When moisture is scarce and control of wind and water erosion necessary, tillage is sometimes carried out in such a way that crop residues are left on the surface. This system is called trash farming, stubble mulch, or subsurface tillage. Principal equipment for subsurface tillage consists of sweeps and rod weeders. Sweeps are V-shaped knives drawn below the surface with cutting planes horizontal. A mounted set of sweeps provided with power lift and depth regulation is often called a field cultivator.

The typical rod weeder consists of a frame with several plowlike beams, each having a bearing at its point. Rods are extended through the bearings, which revolve slowly under power from a drive wheel. The revolving rod runs a few inches below the surface and pulls up vegetative growth; clearance of the growth from the rod is assisted by its rotation. Rod weeders are sometimes attached to chisel plows.

Some control of weeds is obtained by tillage that leaves the middles between crop rows loose and cloddy. When a good seedbed is prepared only in the row, the seeded crop can become established ahead of the weeds. Plowing with the moldboard plow buries the weed seeds, retards their sprouting, and tends to reduce the operations needed to control them. If weed infestations become bad, they can be reduced somewhat by undercutting.

Since rainfall amount and distribution seldom match crop needs, farmers usually prefer tillage methods that encourage soil-moisture storage at times when crops are not growing. From the soil-moisture standpoint, any tillage practice that does not control weeds and result in greater moisture intake and retention during the storage period is probably unnecessary or undesirable.

Minimum tillage. The use of cropping systems with minimal tillage is usually desirable, because intensive tillage tends to break down soil structure. Techniques such as mulching also help prevent raindrops from injuring the surface structure. Excessive tillage leaves the soil susceptible to crusting, impedes water intake, increases runoff, and thus reduces water storage for crop use. Intensive vegetable production in warm climates where three crops per year may be grown on the same land may reduce the soil to a single-grain structure that facilitates surface cementation and poor aeration.

The loosening and granulating actions of plowing may improve soil structure if the plowing is done when the moisture content is optimum; if not so timed, however, plowing can create unfavourable structure. The lifting and inversion of the furrow slice likewise may not always be desirable, because in many cases it is better to leave a trashy surface.

Seeding in narrow slits in sod

The concept of minimum tillage has received much attention. One type of minimum tillage consists in seeding small grain in sod that has been relatively undisturbed. Narrow slits are cut in the sod and seed and fertilizer placed in the breaks thus formed. Soil normally subject to erosion can be planted to grain this way while still retaining the erosion resistance of the sod. The technique has been successful in preparing winter grazing in southeastern portions of the United States. In another type of minimum tillage, the land is broken and planted without further tillage in seedbed preparation. One approach involves breaking the land and planting seeds in the tractor tracks (wheel-track planting); the tractor weight crushes clods and leaves the seed surrounded by firm soil. Another method consists of mounting a planter behind the plow, thus planting without further traffic and leaving a loose seedbed that is satisfactory in areas where postplanting rains may be heavy. In some areas, where winter rain often comes after wheat is drilled, a rotation of wheat following peas has been successful. After the peas have been harvested, the field is rough plowed, and fall wheat is then drilled in directly. All these methods minimize expense and land preparation, tending to leave the soil rough, which reduces erosion and increases water intake. Somewhat similar systems are employed with row crops, where chemical weed control assists in reducing need for cultivation.

Mulch tillage. Mulch tillage has been mentioned already; in this system, crop residues are left on the surface, and subsurface tillage leaves them relatively undisturbed. In dryland areas, a maximum amount of mulch is left on the surface; in more humid regions, however, some of the mulch is buried. Planting is accomplished with disk openers that go through several inches of mulch. Since mulch decomposition may deprive the crop of nitrogen, extra fertilizer is often placed below the mulch in humid areas. In rainy sections, intercropping extends the protection against erosion provided by mulches. Intercrops are typically small grains or sod crops such as alfalfa or clover grown between the rows of a field crop that reach maturity shortly after the field crop has been established and furnish mulch cover for a long time (Figure 9).

Intercropping as protection against erosion

By courtesy of the U S Department of Agriculture

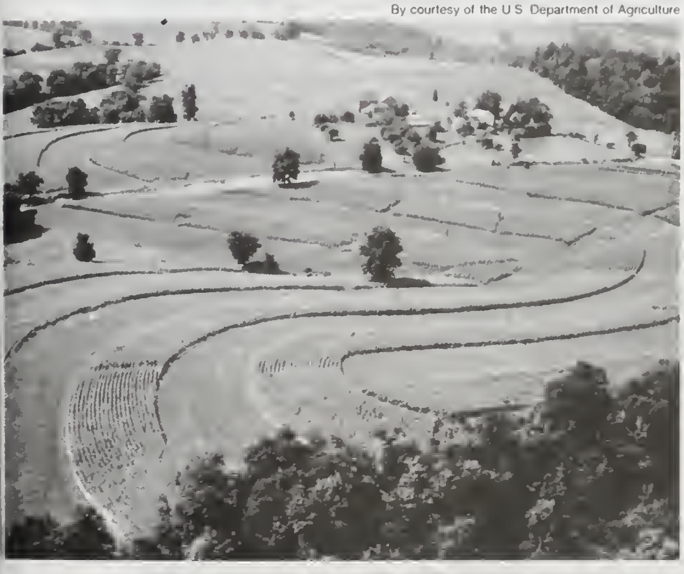


Figure 9: Intercropping to reduce soil erosion, near Fresno, Ohio.

If growth of the intercrop competes with the main crop for moisture and nutrients, that growth may be killed at seeding time or soon thereafter by undercutting with sweeps.

Tillage in dry areas must make maximum use of scanty rainfall. The lister (double-mold board) plow, or middlebreaker, is here used to make water-impounding ridges that promote infiltration. The special problems of dryland farming will be considered below (see *Regional variations: Dryland farming*).

FERTILIZING AND CONDITIONING THE SOIL

Soil fertility is the quality of a soil that enables it to provide compounds in adequate amounts and proper balance to promote growth of plants when other factors (such as light, moisture, temperature, and soil structure) are favourable. Where fertility of a soil is not good, natural or manufactured materials may be added to supply the needed plant nutrients; these are called fertilizers, although the term is generally applied to largely inorganic materials other than lime or gypsum. Fertilizer grade is a conventional expression that indicates the percentage of plant nutrients in a fertilizer; thus, a 10–20–10 grade contains 10 percent nitrogen, 20 percent phosphoric oxide, and 10 percent potash. The green plant, however, requires more nutrients than these.

Essential plant nutrients. In total, the plant has need of at least 16 elements, of which the most important are carbon, hydrogen, oxygen, nitrogen, phosphorus, sulfur, potassium, calcium, and magnesium.

The plant obtains carbon and hydrogen dioxide from the atmosphere; other nutrients are taken up from the soil. Although the plant contains sodium, iodine, and cobalt, these are apparently not essential. This is also true of silicon and aluminum.

Overall chemical analyses indicate that the total supply of nutrients in soils is usually high in comparison with the requirements of crop plants. Much of this potential supply, however, is bound tightly in forms that are not released to crops fast enough to give satisfactory growth. Because of this, the farmer is interested in measuring the available nutrient supply as contrasted to the total quantities. This point will be considered later.

Organic and inorganic soils

The solid content of soils is broadly classified as organic and inorganic. Materials of organic origin range from fresh plant tissue to the more or less stable black or brown degradation product (humus) formed by biological decay. The organic matter is a potential source of nitrogen, phosphorus, and sulfur; it contains more than 95 percent of the total nitrogen, 5 to 60 percent of the total phosphorus, and 10 to 80 percent of the total sulfur. These three elements are cycled through the entire environment of living things (the biosphere). The soil organic matter can be considered as one of the storage points in these cycles. Where nonlegumes are grown in the absence of fertilizer or manures, the crop must gain its nitrogen supply from the organic matter; only a part, however, of the needed phosphorus and sulfur is so supplied.

The inorganic or mineral fraction, which comprises the bulk of most soils, is derived from rocks and their degradation products. The power to supply plant nutrients is much greater in the larger particles, sand and silt, than in the fine particles, or clay. The minerals that comprise the sand and silt in soil contain most of the elements essential for plant growth as a part of their structure. The difficulty is that these minerals decompose so slowly in soil that the rate of supply of the nutrient elements is usually insufficient for good growth of plants.

When the available supply of a given nutrient becomes depleted, its absence becomes a limiting factor in plant growth, and the addition of this nutrient to the soil will increase yields of dry matter. Excessive quantities of some nutrients may cause decrease in yield, however.

Determining nutrient needs. Determination of a crop's nutrient needs is an essential aspect of fertilizer technology. The appearance of a growing crop may indicate need of fertilizer; in some plants, however, the need for more or different nutrients may not be easily observable. If such a problem exists, its nature must be diagnosed, the degree of deficiency must be determined, and the amount and kind of fertilizer needed for a given yield must be found. There is no substitute for detailed examination of plants and soil conditions in the field, followed by simple fertilizer tests, quick tests of plant tissues, and analysis of soils and plants.

Sometimes plants show symptoms of poor nutrition. Chlorosis (general yellow or pale-green colour), for example, indicates lack of sulfur and nitrogen. Iron deficiency produces white or pale-yellow tissue. Symptoms can be misinterpreted, however. Plant disease can produce appearances resembling mineral deficiency, as can various organisms. Drought or improper cultivation or fertilizer application each may create deficiency symptoms.

Strip tests of fertilizer elements

After field diagnosis, the conclusions may be confirmed by experiments in a greenhouse or by making strip tests in the field. In strip tests, the fertilizer elements suspected of being deficient are added, singly or in combination, and the resulting plant growth observed. Next, it is necessary to determine the extent of the deficiency.

An experiment in the field can be conducted by adding nutrients to the crop at various rates. The resulting response of yield in relation to amount of nutrient supplied will indicate the supplying power of the unfertilized soil in terms of bushels or tons of produce. If the increase in yield is large, this practice will show that the soil has too little of a given nutrient. Such field experiments may not be practical, because they can cost too much in time and money. Soil-testing laboratories are available in most areas; they conduct chemical soil tests to estimate the availability of nutrients. Commercial soil-testing kits give results that may be very inaccurate, depending on techniques and interpretation. Actually, the most accurate system consists of laboratory analysis of the nutrient content of plant parts, such as the leaf. The results, when correlated with yield response to fertilizer application in field experiments, can give the best estimate of deficiency. Further development of remote sensing techniques, such as infrared photography, are under study and may ultimately become the most valuable technique for such estimates.

The economics of fertilizers. The practical goal is to determine how much nutrient material to add. Since the farmer wants to know how much profit to expect if he buys extra fertilizer, the tests are interpreted as an estimation of increased crop production that will result from nutrient additions. Cost of nutrients must be balanced against value of crop or even against alternative procedures, such as investing the money in something else with a greater potential return. The law of diminishing returns is well exemplified in fertilizer technology. Past a certain point, equal inputs of chemicals produce less and less yield increase. The goal of the farmer is to use fertilizer in such a way that the most profitable rate is employed.

Fertilizers can aid in making profitable changes in farming. Operators can reduce costs per unit of production and increase the margin of return over total cost by increasing rates of application of fertilizer on principal cash and feed crops. They are then in a position to invest in soil conservation and other improvements that are needed when shifting acreage from surplus crops to other uses.

Farm manure. Among sources of organic matter and plant nutrients, farm manure has been of major importance in past years. Manure is understood to mean the refuse from stables and barnyards, including both excreta and straw or other bedding material, while the term fertilizer refers to chemicals. Large amounts of manure are produced by livestock; such manure has value in maintaining and improving soil because of the plant nutrients, humus, and organic substances contained in it.

As manure must be managed carefully in order to derive the most benefit from it, some farmers may be unwilling to expend the necessary time and effort. Manure must be carefully stored to minimize loss of nutrients, particularly nitrogen. It must be applied to the right kind of crop at the proper time. Also, additional fertilizer may be needed, such as phosphoric oxide, in order to gain full value of the nitrogen and potash that are contained in manure.

Fertilizer value of manure

Manure is fertilizer graded as approximately 0.5–0.25–0.5 (percentages of nitrogen, phosphoric oxide, and potash), with at least two-thirds of the nitrogen in slow-acting forms. Commercial fertilizer equivalent to one ton (900 kilograms) of average manure can be purchased at a fairly low price. Furthermore, the expense of applying 100 pounds (45 kilograms) of 10–5–10 fertilizer is much less than the cost of applying 20 times as much manure. On properly tilled soils, the returns from fertilizer usually will be greater than from an equivalent amount of manure. The application of manure to a crop cannot be controlled as readily as can granulated fertilizer. In general, manure does not provide all the plant nutrients needed and fails to provide any that cannot be supplied by artificial fertilizers. Thus, there is a tendency to discount the value of manure as fertilizer. In underdeveloped countries, however, where artificial fertilizer may be costly or unavailable and where labour is relatively cheap, manure is attractive as a fertilizer.

The main benefits of manure are indirect. It supplies humus, which improves the soil's physical character by increasing its capacity to absorb and store water, by enhancement of aeration, and by favouring the activities of lower organisms. Manure incorporated into the topsoil will help prevent erosion from heavy rain and slow down evaporation of water from the surface. In effect, the value of manure as a mulching material may be greater than is its value as a source of essential plant nutrients.

Green manuring. In reasonably humid areas, the practice of green manuring can improve yield and soil qualities. A green-manure crop is grown and plowed under for its beneficial effects, although during its growth it may be grazed. These green crops are usually annuals, either grasses or legumes, whose roots bear nodule bacteria capable of fixing atmospheric nitrogen. Among the advantages of green-manure crops are the addition of nitrogen to the soil, increase in general fertility level, reduction of erosion, improvement of physical condition, and reduction of nutrient loss from leaching. Disadvantages include the chance of not obtaining a satisfactory growth; the possibility that the cost of growing the manure crop may exceed the cost of applying commercial nitrogen; possible increases in disease, insects, and nematodes (parasitic worms); and possible exhaustion of soil moisture by the crop.

Advantages of green-manure crops

Green-manure crops are usually planted in the fall and turned under in the spring before the summer crop is sown. Their value as a source of nitrogen, particularly that of the legumes, is unquestioned for certain crops such as potatoes, cotton, and corn (maize); for other crops, such as peanuts (groundnuts; themselves legumes), the practice is questionable. Farmers are gradually turning away from growing green-manure crops except where the crop may also serve as winter cover for the land.

Compost, peat, and sludge. Compost, peat, and sludge are used in agriculture and gardening as soil amendments rather than as fertilizers, because they have a low content of plant nutrients. They may be incorporated into the soil or mulched on the surface. Heavy rates of application are common.

Compost, or synthetic manure, is basically a mass of rotted organic matter made from waste-plant residues. Addition of nitrogen during decomposition is usually advisable. The result is a crumbly material that when added to soil does not compete with the crop for nitrogen. When properly prepared, it is free of obnoxious odours. Composts commonly contain about 2 percent nitrogen, 0.5 to 1 percent phosphorus, and about 2 percent potassium; if phosphate and potash are added while composting, those values are higher. The nitrogen of compost becomes available slowly and never approaches that available from inorganic sources. This slow release of nitrogen reduces leaching and extends availability over the whole growing season. Composts are essentially fertilizers with low nutrient content, which explains why large amounts are applied. The maximum benefits of composts on soil structure (better aggregation, pore spacing, and water storage) and on crop yield usually occur after several years of use.

In practical farming, the use of composted plant residues must be compared to the use of fresh residues. More beneficial soil effects usually accrue with less labour by simply turning under fresh residues; also, since one-half the organic matter is lost in composting, fresh residues applied at the same rate will cover twice the area that composted residues would cover. In areas where commercial fertilizers are expensive, labour is cheap, and implements are simple, however, composting meets the needs and is a logical practice.

Peat and peat moss

Peat, composed of prehistoric plant remains that have accumulated under airless conditions in bogs, is a widely used organic soil amendment. Peat moss, the remains of sphagnum plants, is probably its most common form; it contains less than 1 percent nitrogen, with phosphorus and potassium below 0.1 percent. It is highly acid, with pH between 3 and 4.5 (a pH value of 7 is neutral and one above 7 basic). Peat improves the water-storage capability of soils and gives better structure to fine soils. Heavy applications of peat is usually the practice. It is used mostly by specialty-crop producers and on lawns and gardens.

Sewage sludge is the solid material remaining from the treatment of sewage. Its value for soil improvement depends on the method used for treating the sewage. Activated sludge, which results from aerobic (oxygen) treatment, contains 5 to 6 percent nitrogen and 1 to 3.5 percent of phosphorus. After suitable processing, it is sold as fertilizer and soil amendment for use on lawns, parks, and golf courses. It is rarely used in farming.

Liming. Liming to reduce soil acidity is practiced extensively in humid areas where rainfall leaches calcium and magnesium from the soil, thus creating an acid condition. Calcium and magnesium are major plant nutrients supplied by liming materials. Ground limestone is widely used for this purpose; its active agent, calcium carbonate, reacts with the soil to reduce its acidity. The calcium is then available for plant use. The typical limestones, especially dolomitic, contain magnesium carbonate as well, thus also supplying magnesium to the plant.

Another liming material is basic slag, a by-product of steel manufacture; its active ingredient is calcium silicate. Marl and chalk are soft, impure forms of limestone and are sometimes used as liming materials, as are oyster shells. Calcium sulfate (gypsum) and calcium chloride, however, are unsuitable for liming, for, although their calcium is readily soluble, they leave behind a residue that is harmful.

Lime is applied by mixing it uniformly with the surface layer of the soil. It may be applied at any time of the year on land plowed for spring crops or winter grain or on permanent pasture. After application, plowing, disking, or harrowing will mix it with the soil. Such tillage is usually necessary, because calcium migrates slowly downward in most soils. Lime is usually applied by trucks specially equipped and owned by custom operators.

Methods of application. Fertilizers may be added to soil in solid, liquid, or gaseous forms, the choice depending on many factors. Generally, the farmer tries to obtain satisfactory yield at minimum cost in money and labour.

Liquid or solid application of manure

Manure can be applied as a liquid or a solid. When accumulated as a liquid from livestock areas, it may be stored in tanks until needed and then pumped into a distributing machine or into a sprinkler irrigation system. The method reduces labour, but the noxious odours are objectionable. The solid-manure spreader conveys the material to the field, shreds it, and spreads it uniformly over the land. The process can be carried out during convenient times, including winter, but rarely when the crop is growing.

Application of granulated or pelleted solid fertilizer has been aided by improved equipment design. Such devices, depending on design, can deposit fertilizer at the time of planting, side-dress a growing crop, or broadcast the material. Fertilizer attachments are available for most tractor-mounted planters and cultivators and for grain drills and some types of plows. They deposit fertilizer with the seed when planted, without damage to the seed, yet the nutrient is readily available during early growth. Placement of the fertilizer varies according to the types of crops; some crops require banding above the seed, while others are more successful when the fertilizer band is below the seed.

The use of liquid and ammonia fertilizers is growing, particularly of anhydrous ammonia, which is handled as a liquid under pressure but changes to gas when released to atmospheric pressure. Anhydrous ammonia, however, is highly corrosive, inflammable, and rather dangerous if not handled properly; thus, application equipment is quite specialized. Typically, the applicator is a chisel-shaped blade with a pipe mounted on its rear side to conduct the ammonia five to six inches (13 to 15 centimetres) below the surface. Pipes are fed from a pressure tank mounted above. Mixed liquid fertilizers containing nitrogen, phosphorus, and potassium may be applied directly to the surface—by field sprayers where close-growing crops are raised. Large areas can be covered rapidly by use of aircraft, which can distribute both liquid and dry fertilizer.

The future for fertilizers. Future trends in fertilizer technology may be predicted by extrapolating from current developments. Mixtures and materials with high percentages of plant nutrients will dominate the field. Better ways of providing nitrogen, the most expensive of the three major nutrients, will be forthcoming, including increased use of anhydrous ammonia, ammonium nitrate, and urea. Nonleachable nitrogen, for example, can be obtained through the urea–formaldehyde (ureaform) reaction, and ammonium metaphosphate offers a concentrated liquid product. Micronutrients, or trace elements, specific to particular geographical areas will come into increasing use, as will custom mixing and bulk selling of mixtures containing several nutrients based on reliable soil and plant data.

New methods of providing nitrogen

"Complete environment" seeding in which seed, fertilizer, and water are incorporated in a biodegradable (decomposable in the soil) tape may come into use; with the tape planted, no further fertilizer or water will be needed until growth is well established. Such techniques using biodegradable tapes have already been developed on a small scale for use by home gardeners. Finally, larger and more precise fertilizing machines will be developed and adopted.

Factors in cropping

CROPPING SYSTEMS

The kind and sequence of crops grown over a period of time on a given area of soil can be described as the cropping system. It may be a pattern of regular rotation of different crops or one of growing only one crop year after year on the same area.

Crop rotation. Early agricultural experiments showed the value of crop rotations that included a legume sod crop in the regular sequence. Such a system generally maintains productivity, aids in keeping soil structure favourable, and tends to reduce erosion. Alfalfa, sweet clover, red clover, and Ladino clover are considered effective for building up nitrogen. Some legumes, however, do not leave nitrogen behind in the soil because it is deposited as protein in the harvested seed; soybeans are an example. Turning under the top growth of a legume aids in adding nitrogen. Though yields of grains are higher when they are rotated with legumes, it is difficult to determine how much of the improvement depends on the nitrogen added by the legume and how much on improved soil structure or fewer insects and disease.

The determination of the best rotation depends upon whether the crops compete with each other (*i.e.*, if growing one crop lowers the yield of its successor) or complement each other; and the output of one crop on a given acreage leads to increased output of the other. This desirable complementary relationship exists only when one crop or soil-management practice concurrent with it provides nutrient or conditions required by the other crop. In this circumstance, grasses and legumes may complement grains or row crops by furnishing nitrogen, controlling erosion and pests, and improving soil structure to such an extent that greater production is achieved. The reverse can also occur; in certain prairie soils, continuous growing of deep-rooted legumes depletes soil moisture, and subsequent forage yield is improved by frequent plowing of the sod and planting of corn. In high-rainfall or irrigated areas, forage stands deteriorate from winter killing, disease, or grazing, to a point where a year of grain in the rotation allows an improved stand of forage later. Fallow (idle) land is complementary to wheat and other small grains in subhumid areas such as the Great Plains of the United States; such rotation is quite beneficial to wheat yield. Complementary relationships between crops can be terminated by the application of the physical law of diminishing returns, however, and give way to competition.

Rotation as a benefit to wheat yield

Both long-range and short-range profits motivate the farmer as cropping systems are examined in relationship to soil erosion. Excessive loss of soil to streams, rivers, and reservoirs is unacceptable to public policy as well as economically damaging to the farmer, and crop rotations that promote erosion are minimized. Soil losses are least from fields in continuous sod and most from continuous row crops. If row crops are grown in rotation with sod, the erosive susceptibility of row crops is reduced over a period of time. Peanuts (groundnuts), potatoes, tobacco, cotton, sugar beets, and some vegetables, and similar row crops that require frequent cultivation (intertillage) and leave minimal post-harvest residue are most likely to permit serious erosion. Less erosive are row crops such as corn (maize), sugarcane, and grain sorghum, which require less cultivation and leave more residue. Small grains such as wheat, oats, barley, and rye usually permit less erosion than the row crops. Among sod crops, grasses or grass–legume mixtures are less erosive than pure stands of legumes such as alfalfa. Fortunately, cropping systems that tend to control soil erosion usually tend also to give better yields than systems that promote excessive erosion. This results from increased availability of water to the plants and increased amounts of nutrients, which in erosive systems are washed away and lost.

Monoculture. The practice of growing the same crop each year on a given acreage, monoculture, has not been generally successful in the past, because nonlegume crops usually exhaust the nitrogen in the soil, with a resulting reduction in yields; this is particularly true in humid regions. The advent of low-cost nitrogen fertilizers, however, has induced reconsideration of the possible advantages of monoculture. These advantages can best be discussed in terms of a hypothetical general farm where it may be desirable to produce several different kinds of crops: the question to be answered is whether monoculture can do better than rotational systems in producing these crops while still maintaining productivity.

Advantages of monoculture. First, if different kinds of soil exist on the farm, a monoculture system may permit each crop to be grown on the soil best suited to it. Forage crops, for example, could be confined to steep land to minimize erosion; intertilled crops could be planted on the better soils with gentle slopes. Wet areas could be used continuously for crops not requiring early-spring field operations, while dry soils could be used for drought-resistant crops such as sorghums or winter small grains.

Second, the fertility level of the soil can be adjusted to fit one crop more precisely than it can be adjusted to fit all the crops in a rotation.

Third, where continuous cropping is practiced and perennial forage crops are used, regular reseedings are avoided. This is an advantage, because each seeding is accompanied by the possibility of failure.

Flexibility in monoculture crop planning

Fourth, systems based on monoculture usually offer greater flexibility in planning the system to meet year to year changes in the need for various crops. Part of the acreage can be shifted from one crop to another without upsetting the total farm cropping plan.

Disadvantages of monoculture. On the other hand, requirements for successful monoculture are more demanding of management skill than are sod-based rotations. The entire nitrogen need of nonlegume crops must be met by purchased fertilizers or by use of manure. Closer attention to soil erosion is necessary, except for perennial sod. Soil-structure problems can become severe where intertilled crops are grown continuously. In monoculture, the farmer is completely dependent on chemical insecticides, disease-resistant plant varieties, soil fumigation, and similar methods of controlling insects and diseases that are usually controlled by crop rotation.

Thus, the choices of cropping systems that maintain good productivity, minimize soil losses, and are in harmony with demand and desired business organization are not easily made. The growing use of systems analysis will undoubtedly aid in rational selection among the bewildering array of possibilities.

CROP PROTECTION

Crops are vulnerable to attack, damage, and competition. Insects, plant disease, nematodes, rodents, weeds, and air pollution are among the many enemies that can reduce crop yields and deny man the use of some of his farm-stored crops.

Insects, for example, can destroy a crop in a relatively short time. Control measures for many years have engaged the attention of farmer and scientist, yet full success has not been achieved, and the battle continues. The problem is further complicated by the fact that control measures not only kill unwanted insects, but also may harm honey bees as well as the parasites and predators that destroy insect pests.

At least 10,000 species of insects are classed as unwanted. Of these, several hundred species are particularly destructive and require some degree of control. They destroy food as well as the forage, pasture, and grain needed to produce livestock; and, in addition, they carry and transmit many diseases of plants and animals.

Adverse effects of some insecticides

Chemical control of insects. Insecticides generally are effective, cheap, and safe if handled correctly; the good derived from them, however, can be partly offset by adverse effects. Chlorinated hydrocarbon insecticides such as DDT, for example, may leave residues toxic to beneficial insects, fish, and other wildlife; the insecticides may be found in meat and milk, or they may persist in the soil. Another problem is that some species of insects build up resistance to chlorinated hydrocarbon, organic phosphate, and carbamate insecticides. These disadvantages can be overcome only by persistent search for new and safer insecticides accompanied by wide use of nonchemical insect control.

A wide range of organophosphate and carbamate materials is now available. These can be applied to avoid most of the problems related to residues. Malathion and carbaryl, for example, are used to control insects in areas where persistent materials might appear later in meat or milk and can also be applied in areas where fish and wildlife might be affected. Those two chemicals offer a broad range of toxicity to insect pests. Unlike chlorinated hydrocarbons, they can be applied up to within a day or so of harvest without harm to many crops; they are dangerous, however, to those who apply them and must be handled with care.

Some insecticides are effective in very small amounts. This fact has stimulated development of ultralow-volume technology, where special equipment permits dispersal of low volumes of undiluted chemicals, which offers cost advantages as well as drastic reduction of the chemicals in the environment. For example, six to 16 ounces (170 to 450 grams) per acre of Malathion may be effective against grasshoppers, boll weevil, cereal-leaf beetle (*Oulema melanopus*), mosquitoes, and the beet leafhopper (*Circulifer tenellus*). Formulation of chemicals in granules rather than sprays offers some advantages in use and applications; among others, it reduces the amount needed and also lessens the chance of adverse effects on beneficial insects and wildlife.

Systemic insecticides

Certain insects that attack cotton, vegetables, and forage crops may be controlled by chemicals absorbed by the plant. Called systemics, they are placed with the seed at planting time. The chemical is taken up by the plant, and insects die when they attempt to feed on the leaf or stem. Beneficial insects that do not feed on the plant remain unharmed.

Nonchemical control of insects. *Mechanical and cultural controls.* Light traps that give off radiation that attracts insects have been under test for many years. They have been somewhat successful in controlling the codling moth (*Carpocapsa pomonella*) and the tobacco hornworm (*Protoparce sexta*).

Use of reflective aluminum strips, placed like a mulch in vegetable fields, has reduced or prevented aphid attack and thus protected cucumbers, squash, and watermelons from mosaic diseases. This technique may supplant insecticides, which frequently do not kill aphids quickly enough to prevent crop losses from virus transmitted by them.

For stored products, heat or cold can control many insects that frequent such places. Also, changing the proportions of oxygen, nitrogen, and carbon dioxide in the storage atmosphere can provide control.

Recently, it was discovered that, if adult Indian-meal moths (*Plodia interpunctella*) were exposed to certain wavelengths of sound during the egg-laying period, their reproduction was reduced by 75 percent. The sound waves had a similar effect on flour beetles (*Tribolium* species). Further development is needed, but this method offers potential as a nonchemical control. Other types of physical energy can also kill insects. Light waves, high-frequency electric fields, high-intensity radio frequencies, and gamma radiation have been investigated; some offer promise.

Certain cultural practices can prevent or reduce insect crop damage. These include destruction of crop residues, deep plowing, crop rotation, use of fertilizers, strip-cropping, irrigation, and scheduled planting operations. Such practices are useful but cannot be relied upon entirely to eliminate severe infestations.

Biological controls. The question of using biological controls has always been of considerable public interest. The control agents include parasites, predators, diseases, protozoa, and nematodes that attack the insect pests. Biological controls cannot replace insecticides entirely, because nature provides for survival of both beneficial and destructive insects. Before the population of a parasite or predator can expand, a high population of the host species must also be present. Sometimes the control agents are far outnumbered by the pest insect. Parasites and predators have funished good control of the Japanese beetle (*Popillia japonica*), European corn borer (*Pyrausta nubilalis*), alfalfa aphid (*Therioaphis maculata*), alfalfa weevil (*Hypera postica*), and several others.

Microbial agents as controls

Microbial agents can be used for control. There exist about 1,100 viruses, bacteria, fungi, protozoa, rickettsiae, and nematodes that parasitize insects. Many pathogens are specific to a particular insect but are harmless to man and domestic animals. It is a possibility that insect pathogens can be produced, packaged, distributed, and applied in much the same way as insecticides.

The ideal solution to insect-control problems is to plant crop varieties that are resistant to attack. The only difficulty is that such varieties are not universally available, and development entails a very long process.

Sterilization of male insects by gamma radiation and their release into a population of wild insects is a promising approach. It has proved successful in control of screwworms and fruit flies, replacing chemicals in some areas. Chemical attractants, which lure insects into contact with small amounts of insecticide or a sterilant, also offer much promise.

All aspects of insect control considered, it is possible that "integrated control," coordinated employment of more than one method, may be the answer. Combining resistant varieties with a systemic insecticide that leaves the parasites and predators unharmed, for example, has been successful in combatting the spotted alfalfa aphid in California. Preliminary reduction of heavy infestation by chemical spray combined with bait, followed by the sterile-insect technique, provides another example of integrated control. Use of sex attractant in light traps, plus special management of postharvest residues, has controlled the tobacco hornworm. Other examples might be cited, but the principal value of such control methods lies in using less insecticide and thus contributing to maintenance of a good environment.

Control of plant diseases and nematodes. Insects, of course, are not the only agents hazardous to crops. Plant diseases and the microscopic worms called nematodes have the potential of creating wholesale destruction of crops, especially those grown in regions of wide weather fluctuation. In fact, these plant pests sometimes limit the kinds and varieties of crops that can be grown. The damage they cause may sometimes be mistaken for that caused by unfavourable weather. Epidemics may destroy crops completely.

As with insects, control of plant diseases and nematodes covers a broad spectrum of measures: use of chemicals, resistant varieties, quarantine, forecasting and warning, cultural practices, heat treatment, and others. Furthermore, most plant virus diseases are transmitted by insect carriers, so control of insects is linked to control of disease.

Nematodes and plant disease can at times be controlled fairly well by crop rotation, deep plowing, and burning of stubble and debris that remain after harvest. Though burning destroys aboveground organisms and permits economical control by chemicals, it contributes to air pollution and destroys organic matter. In another technique, propane-gas flame is applied to living plants as well as stubble to kill disease spores. A virus disease of sugarcane is controlled by heating diseased cuttings in hot-air ovens. Stem rot disease of peanuts (groundnuts) can be controlled by plowing under dead plant debris or by planting the seed on a raised bed followed by application of a pre-emergence weed killer.

Predicting plant epidemics

Successful control of epidemic plant disease may depend on prompt application of chemicals before the disease outbreak. Many governments operate plant-disease forecasting and warning services for farmers. The service is based primarily on analysis of temperature, rainfall, humidity, and dew—all factors that can create conditions favourable to disease outbreaks.

Weed control. Weed control is vital to agriculture, because weeds decrease yields, increase production costs, interfere with harvest, and lower product quality. Weeds also impede irrigation water-flow, interfere with pesticide application, and harbour disease organisms.

Early methods of weed control included mowing, flooding, cultivating, smothering, burning, and crop rotation. Though these methods are still important, other means are perhaps more typical today, particularly the use of herbicide (plant-killing) chemicals. Another technique is to introduce insects that attack only the unwanted plant and destroy it while leaving the crop plants unharmed.

The inadequacy of the cultural, mechanical, and biological control systems, however, stimulated the rapid development of chemical usage since World War II. Herbicides have had an impact on crop production, changing many cultural and mechanical agricultural operations.

Herbicides are formulated as wettable powders, granular materials, emulsions, and solutions. Any of them may be applied as a spot treatment, broadcast, placed in bands, or put directly on a specific plant part. When formulated as solutions or emulsions, the chemical is mixed with water or oil.

Spraying is the most common method, permitting extremely small amounts to be applied uniformly because of dilution. Sprays can be accurately directed underneath growing plants, and calibration and rate control are easier with spray machines than with granular applicators. Granular formulations have advantages under some conditions, however. The use of herbicides must be integrated into the overall farm program because the optimum date and application rate depend on the crop stage, the weed stage, weather conditions, and other factors.

Careful use of herbicides in farm production lowers cost, resulting in a more economical product for the consumer. Herbicides cut the costs of raising cotton, for example, by reducing labour requirements for weed control up to 60 percent. Herbicides replace hand labour in growing crops, labour that is no longer available in developed nations at costs the farmer can afford. Machines for chemical application are widely available.

Safety of herbicides

When used as directed, herbicides are generally safe, not only for the operator but also for wildlife and livestock.

The greatest difficulty lies in accidental injury to crop plants resulting from drift and from residues in the soil, particularily if residues enter water courses.

The future of chemical pesticides and herbicides is under debate by those who manufacture, sell, and use them and by those who are concerned about environmental quality. The value of an assured food and fibre supply at reasonable cost is undeniable, and chemicals contribute much toward this. These substances also cause undesirable effects upon the environment, however, and indeed can be toxic to a wide range of organisms. This fact will demand an increasing amount of care in using chemicals, perhaps enforced by law, along with increasing use of nonchemical control techniques.

Harvesting and crop processing

HARVESTING MACHINERY

Reaping and threshing processes

Harvesting machinery is generally classified by crop: reapers for cutting cereal grains and threshers for separating the seed from the plant. The more modern combine cuts, threshes, and cleans the grain in one operation. Corn (maize) harvesting is performed by mechanical corn pickers that snap the ears from the stalk so that only the grain and cobs are harvested. Corn shelling may be done mechanically in the field, after or with picking. Stripper-type cotton harvesters, which strip the entire plant of both open and unopened bolls, work best late in the season after frost has killed the green vegetative growth. Hay and forage machines include mowers, crushers, windrowers, field choppers, balers, and some machines that press the hay into wafers or pellets.

Grass, legumes, corn (maize), and other crops are often put into silos to keep them in a succulent and fermented state rather than stored dry as hay. To make silage, the crops must be cut up to permit tight packing in the silo, producing anaerobic fermentation and preventing formation of mold. Almost all silage crops are cut in the field with a forage harvester that cuts and chops the crop immediately or picks up and chops a windrow that has been cut and raked earlier.

Root crops are harvested with diggers and digger-pickers, which often pull up clods, stones, and vines with the crop. Though some machines carry workers who manually sort out extraneous material, this task is increasingly being performed mechanically. Modern sugar-beet harvesters lift the whole root from the ground, clean the earth from it, and deliver it to a bin or wagon. Sometimes the beet tops are removed before harvest of the roots and used for cattle feed. Peanuts (groundnuts) are lifted, vines and all, and allowed to dry before removal of the pods.

Three types of tobacco harvest methods

Tobacco-harvesting aids may be classified in three principal ways, according to the harvesting and curing methods used, which depend on the type of tobacco and its use. Flue-cured tobacco, a large plant that may stand three to four feet (90 to 120 centimetres) high, is harvested with machines that carry several workers who ride the lower platforms of the machines, cut the leaves, and place them on conveyor belts, where the leaves are tied mechanically or by hand. Burley tobacco has usually been harvested by workers using a machete-type knife. After cutting, the large end of the stalk is fixed onto the sharpened end of a stick, which—when loaded with a number of stalks—is hung by hand in a tobacco barn for curing. Researchers are attempting to mechanize the cutting, impaling, and hanging of burley tobacco. Little has been done, however, toward the mechanization of the harvesting of the small aromatic tobacco leaves, which are grown in the shade, picked by hand, tied with a string, then hung for curing.

Tree-crop harvesting, accomplished by hand or with mechanical shakers, is described below under *Fruit farming*. Vegetable crops such as asparagus, lettuce, and cabbage are still harvested largely by hand, though scarcity and high cost of field labour has led to some mechanization in this area, notably with tomatoes.

CROP-PROCESSING MACHINERY

Machinery is widely used to prepare crops for convenient transportation, for safe storage, for the market, and for feeding to livestock. Advances in such machines have been rapid, particularly with new crops, increased yields, multiple-crop practices, and changing techniques.

In the most common method of crop drying, the crop, usually grain, is spread on floors or mats and stirred frequently while exposed to the sun. Such systems, though extremely common in the underdeveloped countries, are very slow and dependent on the weather. Forced-air-drying systems allow the farmer much more freedom in choosing grain varieties and harvest time. Fairly simple in operation, these systems have been gaining popularity in the tropics. Heat is often added to increase air temperatures during the drying period.

Dryeration of corn

In a process called dryeration, wet corn (maize) is placed in a batch or continuous dryer. After losing 10 to 12 percent of its moisture, the hot corn is transferred to the dryeration cooling bin, in which it is tempered for six to 10 hours and then slowly cooled by ventilation for 10 hours. This process reduces kernel damage and increases dryer output.

High moisture in stored hay not only causes rapid deterioration of its value as feed but often results in spontaneous combustion. When hay is first cut, it usually contains 70 percent or more moisture. It wilts and quickly dries to a moisture content of about 40 percent. At this stage, it can be dried to a safe storage condition, about 15 percent moisture, by blowing air through it, sometimes with supplemental heat.

Feed-processing mills, often referred to as feed grinders, are used principally for milling cereals for livestock feed, which aids digestion. The ground material is usually fairly coarse and at times may only be crushed. Modern mills frequently are designed to allow the farmer to grind the grain and to mix in various other ingredients in desired quantities.

Other types of crop-processing machinery include machines that separate desirable seed from weed seed, stems and leaves, and dirt; grading machinery to classify seed by width, length, or thickness; fruit graders and separators; and cotton gins, which separate cotton seeds from the fibres.

Regional variations in technique

DRYLAND FARMING

Dryland-farming crops

Dryland farming refers to production of crops without irrigation in regions where annual precipitation is less than 20 inches (500 millimetres). Where rainfall is less than 15 inches (400 millimetres) per year, winter wheat is the most favoured crop, although spring wheat is planted in some areas where severe winter killing may occur. (Grain sorghum is another crop grown in these areas.) Where some summer rainfall occurs, dry beans are an important crop. All dryland crop yield is mainly dependent on precipitation, but practices of soil management exert great influence on moisture availability and nutrient supply.

Where rainfall exceeds 15 inches (380 millimetres), the variety of crop possibilities is increased. In areas of favourable soils and moisture, seed alfalfa is grown, as is barley. Some grass seed may be grown, particularly crested wheat grass of various types.

Fallow system and tillage techniques. Dryland farming is made possible mainly by the fallow system of farming, a practice dating from ancient times. Basically, the term fallow refers to land that is plowed and tilled but left unseeded during a growing season. The practice of alternating wheat and fallow assumes that by clean cultivation the moisture received during the fallow period is stored for use during the crop season. Available soil nitrogen increases and weeds are controlled during the fallow period. One risk lies in the exposure of soil while fallow, leaving it susceptible to wind and water erosion. Modern power machinery has tended to reduce this risk.

Procedures and kinds of tillage that are comparatively new have proved effective in controlling erosion and improving water intake. Moldboard and disk plows are being replaced with chisels, sweeps, and other tools that stir and loosen the soil but leave the straw on the surface. Where the amount of straw or residue remaining from the previous crop is not excessive, this trashy fallow system works well, and tillage implements are designed to increase its effectiveness.

Contour tillage helps to prevent excessive runoff on moderate slopes. Broad terraces can aid in such moisture conservation. Steeper slopes are planted to permanent cover.

Improving storage of moisture

Compacted zones at a depth of five to eight inches (13 to 20 centimetres) can be caused by tillage. As such zones interfere with storage of moisture, they can be controlled by growing deep-rooted alfalfa at intervals, or the compacted zone can be broken by fall tillage with chisels or sweeps set to a depth just below the zone of compaction. Such deep tillage will result in reduced runoff and deeper moisture penetration.

When using power machinery in dryland farming, the timing of operations is important. The soil is broken in the fall or early spring before weeds or volunteer grain can deplete the moisture. Use of a rod weeder or similar equipment during fallow can control the weeds. Planting is timed to occur during the short period in fall or spring when temperature and moisture are favourable.

Fertilizer use. Fertilizer is an important component of dryland technology. For example, 20 pounds per acre (22 kilograms per hectare) of nitrogen are recommended where rainfall is less than 13 inches (330 millimetres), ranging up to 60 pounds per acre (67 kilograms per hectare) where more rain is available; those figures refer to the production of wheat, but they are applicable to other dryland-farming areas. Where average annual precipitation is less than 12 inches (300 millimetres), the use of nitrogen is limited to years where moisture outlook is exceptionally favourable. Nitrogen fertilizer can be applied either in fall or spring. Band placement or broadcast techniques are utilized. Good results are obtained from broadcast spring application of nitrate fertilizer, and fall application of ammonia has also been successful. Local climates and rainfall patterns also determine choice of fertilizer and time of application.

Crops and planting methods. Alfalfa grown for seed on drylands is planted in rows, usually two to three feet (60 to 90 centimetres) apart; cultivation between rows is required during the first year. Alfalfa is also grown for forage where favourable. This practice builds nitrogen and organic matter, while improving soil structure. These legumes can be rotated with wheat if rain is between 16 and 18 inches (400 and 450 millimetres) and will increase the yield of wheat.

Other crops, such as cotton, peanuts (groundnuts), and grain sorghum, can be grown successfully in dryland agriculture. Where those crops are produced on sandy soils, special techniques are necessary to reduce soil blowing and drifting. Cotton and peanuts do not produce sufficient crop residues for protection from wind erosion, while sorghum does. For this reason, many farmers in such areas use various row combinations of cotton or peanuts with grain sorghum; two rows of grain sorghum and four to eight rows of peanuts in alternating strips is a popular technique. Another is to use a two-year rotation of cotton and grain sorghum, in which two rows are cropped and two rows are fallow. These systems not only afford protection from wind erosion but also promote effective use of soil moisture.

TROPICAL FARMING

The area of the world bounded roughly on the north by the Tropic of Cancer and on the south by the Tropic of Capricorn, a vast land that embraces large parts of Latin America, Africa, India, Australia, and Southeast Asia, contains climates less favourable to agriculture and human settlement than those of the temperate zones. Within this Equator-centred area occur the climates known as tropical, which are characterized by two general types: warm and wet, and warm with partly deficient rainfall. In either, the total precipitation is usually quite heavy, which leaches the tropical soils of nutrients. The area also has high temperatures with little variation the year round. The combination of high temperature and high rainfall causes organic matter to decompose quickly, leaving the soil deficient in humus. Vegetation flourishes in the tropics, along with weeds, insects, and disease organisms. Important climatic variations occur, depending upon land elevation.

Characteristics of tropical areas

Tropical crops include coconut, palm oil, rice, sugar, pineapple, sisal, cocoa, tea, coffee, jute, rubber, pepper, banana, and many others. In certain highland tropical areas, however, the crops common to temperate-climate agriculture can also be grown. The amount of tropical land well-suited to agriculture, however, is limited.

Plant-pest problems. The abundance of plant pests in the tropics, including weeds and disease, makes agriculture successful mainly in the plantation system, where needed control measures can be financed. The alternative is to move from deteriorated land to newer fields; this practice of shifting agriculture has also been common, because tropical soils lose their productive capacity so rapidly. The practice probably cannot be continued indefinitely, however, because of increasing population pressure.

The largest quantities of commercial tropical products originate in plantations, where skilled management is combined with sufficient capital to provide mechanized equipment. This is particularly true in the production of coffee, cocoa, rubber, coconut, banana, pineapple, sugarcane, and others. Much rice is produced in the Asian tropics and Indonesia, however, on small farms with intensive hand labour and simple tools, where the prime mover is likely to be the ox or the water buffalo, not the tractor.

Guyana's system of coastal water control

Water management. Drainage, irrigation, and other special techniques of water management are important in tropical agriculture. An example is the cultivation of rice and sugarcane in the fertile coastal areas of Guyana. Originally through private enterprise and later by government efforts, large coastal areas were "empoldered" (diked) to keep back the sea in front and floods from the rivers in the rear. With a mean annual rainfall of 90 inches (2,300 millimetres), drainage is a critical factor; in fact, the system cannot discharge all possible floodwater, and so the crops must tolerate occasional drowning. With gravity drainage effective only at low tide, the drainage gates are opened on the ebbing tide and closed on the rising tide. Great difficulty is encountered in keeping the outlets unclogged by the heavy sediment discharge. Since rain does not always fall when it is needed, many fields are irrigated. Most of the rice soil is specially tilled after plowing in order to create a better seedbed under the water, using tractors operating in water four to six inches (10 to 15 centimetres) deep. After this special tillage, the seeds are broadcast in one to two inches (2.5 to five centimetres) of water. Though maintenance and operation of such an intricate water-control system are not simple, Guyana rice production has been doubled through its use.

Mechanical problems. Mechanization faces many obstacles before wide adoption is possible in tropical regions. Difficult soils, stones, stumps, abundant labour, resistance from farmers, lack of incentives, lack of skills, lack of capital, low wages, high cost of machines, lack of dealer service, fragmented land ownership, all contribute to slow development of mechanization. Tropical soils differ markedly from those in the countries that manufacture land-preparation machinery, making adaptation of new design necessary. The encountering of stones, wood, trash, and termite mounds causes machines to break down. Depressing climatic conditions reduce the performance of the machine operators. Tropical farm regions are notoriously irregular or mountainous, impeding intensive machine culture. The best soils in Brazil require special erosion controls, reducing the potential for large-scale mechanization. One of the greatest overall impediments to mechanization is the fear that unemployment might result from it, a failure to understand that economic development and higher living standards depend partly on increasing the productivity of labour.

Mechanization in Trinidad

As an example of the problems encountered in mechanizing tropical crops, the harvesting experience of a large sugarcane plantation in Trinidad is illuminating. On flatland of some 30,000 acres (12,000 hectares), the cane is grown on heavy clay soil in a climate with 50 inches (1,300 millimetres) of rain during the seven-month wet season and 10 inches during the five-month dry season. By 1960 the rising wage rate made harvest mechanization imperative. First, the traditional "bed" system, which functioned to remove floodwater, was changed to ridge planting; this made it possible for machines to operate and was a remarkable change in itself. Then it was decided to harvest with the cane combine, which tops, cuts, chops, and loads the chopped cane into transport vehicles. Although the combine is complicated and requires considerable power, it was deemed better than mechanical half-measures.

By 1969 the combines, however, were harvesting only 12.8 percent of the flatland crop, indicating that mechanization was far from complete. Three factors were responsible: first, the cane combines required extensive maintenance plus very expensive replacement parts. Second, it was difficult to mobilize a transport system to receive the output of the combines with any degree of economy. Third, the social problem of displaced workers had to be considered. The combines increased labour productivity sixfold over hand harvesting; thus, their introduction had to be slowed until surplus workers could be accommodated elsewhere. The limited success of this mechanization project indicates how complicated such a process really is.

Taking the largest view of possibilities for improving tropical agriculture, the most promising inputs of technology are improved crop varieties and increased use of fertilizers.

OTHER SPECIALIZED TECHNIQUES

Hydroponics. The term hydroponics denotes soilless culture of plants. The possibilities of this technique have received considerable attention in recent years. In hydroponics, an outgrowth of laboratory techniques long used by scientists, plants are grown with their roots immersed in a water solution containing necessary minerals or rooted in a sand medium kept moistened by such a solution. Soilless culture of plants is similar in principle but larger in scale. A typical hydroponics technique has plants supported in a bed of peat, wood fibre, or similar material, on a wire screen with the roots dipping into the solution below. Aeration of the solution is provided. In another method, the plants are rooted in a medium of sand or gravel contained in a shallow tank into which the solution is pumped at intervals by automatic control. Between pumpings, the solution drains slowly down into a reservoir tank. Hydroponic techniques are practiced on a small scale both out-of-doors and in greenhouses.

Of the elements known to be necessary for plant growth, carbon, oxygen, and hydrogen are obtained by the plant from atmospheric gases or from soil water. The others are all obtained as mineral salts from the soil. The elements absorbed as salts—iron, manganese, boron, copper, zinc, and molybdenum—are required in minute quantities and are called the micronutrients. The principal elements that must be provided as dissolved salts in hydroponic techniques are nitrogen, phosphorus, sulfur, potassium, calcium, and magnesium. Numerous solutions have been devised to fulfill these requirements.

Crop yields of some plants can be obtained fully equal to those obtained on fertile soils. Wide-scale crop production by hydroponics, however, would be economic only for certain intensive types of agriculture or under special conditions. Some greenhouse crops, both vegetables and flowers, are grown by this method. In regions having no soil or extemely infertile soil but with favourable climate, hydroponic techniques have been very useful; for example, on some of the coral islands of the Pacific.

Greenhouses. The greenhouse is typically a structure whose roof and sides are transparent or translucent, permitting a sufficient quality and quantity of solar radiation to enter the structure for photosynthesis (see below). It allows the growing of crops independently of the outside climate, since its interior temperature and humidity can be controlled. Greenhouses vary in size and complexity from small home or hobby structures to large commercial units covering an acre or more of land. An even smaller greenhouse might be termed the hot bed, a glass-topped box containing fermenting organic matter; the fermentation process yields heat, allowing the gardener to start plants from seed in early spring for later transplanting.

The basic construction of a greenhouse consists of a light but sturdy frame capable of resisting winds and other loads. Conventional foundations usually support vertical walls; the roof may be gabled, trussed, or arched. The conventional greenhouse is fitted with glass panes, but plastic-film or fibre-glass panels often supplant glass.

Basic greenhouse construction

Maintenance of temperature within the greenhouse is difficult because of fluctuating outside conditions. When the sun shines brightly, little heat is needed, and the heating system must be controlled in some way to prevent injury to the crop. Hot water, steam, electric cable, or warm-air furnaces provide the heat, which is usually controlled by thermostat. Temperatures in greenhouses are regulated to suit the crop. Typical ranges are from 40° F (4° C) for lettuce, violets, carnations, and sweet peas to 70° F (21° C) for cucumbers, tomatoes, and orchids.

Cooling is often required during summer days in warm climates. Ventilation is the simplest technique, reducing inside temperature to near that of the outdoors. Additional cooling by refrigeration may be required; in dry regions, the evaporative cooler is efficient and also increases the relative humidity within the structure. Another form of environmental control consists of adding extra carbon dioxide to the air if the crop requires it for extra photosynthetic efficiency.

The commercial-greenhouse operator usually grows vegetables or ornamental plants. Such production makes more demands on the grower, because he must assume many of the tasks normally handled by nature in the open fields. He must regulate the temperature, ventilate, adjust the amount of entering sunlight, provide soil moisture, fertilize, and even facilitate pollination. During the off-season, the structure must be cleaned and fumigated, its soil restructured, and mechanical equipment checked. Mechanization of greenhouse operations has lagged far behind the pattern of agriculture in general. Disease is a particularly serious hazard in greenhouse farming, requiring constant attention and use of chemicals.

The factor of weather

WEATHER INFORMATION

The interaction of weather and living systems is a basic aspect of agriculture. Although great strides in technology have resulted in massive production increases and improved quality, weather remains an important limiting factor. Though man is not yet able to change the weather, except on a very small scale, he is capable of adjusting agricultural practices to fit the climate. Thus, weather information is of utmost importance when combined with other factors, such as knowledge of crop or livestock response to weather factors; the farmer's capability to act on alternative decisions based on available weather information; existence of two-way communication by which specific weather forecasts and allied information can be requested and distributed; and the climatic probability of occurrence of influential weather elements and the ability of the meteorologist to predict their occurrence.

Other weather-research benefits. Apart from the many applications of weather forecasting to current problems, meteorological research may benefit agriculture in at least three other ways: (1) improved planning of widescale land usage depends partly on detailed knowledge of plant-climate interactions; radiation, evapotranspiration, diurnal temperature range, water balance, and other parameters are measured and analyzed before a plan realizing maximum economic benefit for a given area is prepared; (2) agronomic experiments are combined with climatological documentation to obtain the greatest scientific and technological return; (3) problems of irrigation, row spacing, timing of fertilizer application, variety selection, and transplanting can best be solved with the aid of climatic environmental data; cultural practices related to artificial modification of microclimates should be based on research knowledge rather than personal judgment.

Observing climatic elements. The climatic elements the observation of which is valuable for agricultural purposes can be approached on an idealized threefold scale: (1) microscale observations of small areas for research designed to elucidate basic physical processes; (2) mesoscale climatic networks designed for practicing farmers to improve

Three classes of weather stations

their operations; and (3) macroscale regional networks intended for weather forecasting and for gathering basic climatic data (see also CLIMATE AND WEATHER). Macroscale stations can be further divided into first-order and second-order stations, the number and type of observations different for each. Micrometeorology demands the most elaborate array of measuring devices, while a second-order macroscale station requires the least; in fact, the latter station will measure only five elements: air temperature, rain, snow, humidity, and surface wind. A first-order macrostation will be equipped to measure 16 elements: global radiation, sunshine hours, clouds, net radiation, air temperature, soil temperature, rain, snow, hail, dew, fog, humidity, pan evaporation, pressure, upper air wind, and surface wind. Mesoscale measurements include 10 elements and microscale 27 (three of which are derived from others).

The World Meteorological Organization and the various national weather services are concerned with establishment and improvement of macroscale regional climatic stations, both first-class and second-class. Spaced at least 10 miles (16 kilometres) apart, their value for daily agricultural operations is limited, but they are useful for long-range planning and forecasting. Most parts of North America, Europe, and Australia have adequate networks of these stations, but wide gaps exist in the tropics, polar regions, and arid lands.

The degree day. One weather characteristic of agricultural value is the degree day. This concept holds that the growth of a plant is dependent on the total amount of heat to which it is subjected during its lifetime, accumulated as degree days. Common practice is to use 50° F (10° C) as a base. Thus, if the mean daily temperature for a particular day is 60° F (16° C), then 10 degree days are accumulated for that day on the Fahrenheit scale. The total number of growing degree days required for maturity varies with crop variety as well as plant species. Also, the minimum threshold temperature (the temperature below which the plant is damaged or unable to grow) varies with plants; *e.g.*, 40° F (4° C) for peas, 50° F (10° C) for corn (maize), and 55° F (13° C) for citrus fruits. Where studies have established the number of degree days required for maturity of a given crop, the planting dates can be scheduled for orderly harvest and processing. The system is helpful in selecting crop varieties appropriate to different geographical areas; it also has value in scheduling spray programs and predicting insect emergence.

Weaknesses of the degree-day concept

The growing-degree-day concept has certain weaknesses: (1) it assumes that the relationship between growth and temperature is linear (actually it is not); (2) it makes no allowance for changing threshold temperatures with advancing crop development; (3) too much weight is given to temperatures above 80° F (27° C), which may be detrimental; and (4) no account is taken of the diurnal temperature range, which is often more significant than the mean daily value.

WEATHER EFFECTS

The essence of the weather–agriculture interaction for the farmer lies in wise adaptation of operations to the local climate and in techniques for manipulating or modifying the local environment (microclimate) to minimize weather stresses on plants and animals. Many of these techniques have been practiced for centuries: seeding and cultivation, irrigation, frost protection, animal shelters, windbreaks, and others are methods of altering the microclimate. The climatic factors and their relation to plant growth in terms of protective techniques are important.

Solar radiation. Solar radiation is the ultimate source for all physical and biological processes of the earth. Agriculture itself is a strategy for exploitation of solar energy, made possible by water and nutrients. During daytime hours, solar radiation is delivered both directly and by diffused sky reflection. The incoming radiation that is not reflected by the surface or reradiated to outer space is the net radiation, which is the energy available for maintaining the earth's surface temperature. At night the net radiation is negative; that is, energy is lost to outer space by long-wave radiation, and none is gained. The net radiation balance varies widely throughout the world, setting limits on basic agricultural possibilities.

Leaf canopy's effect on solar-energy distribution

Photosynthesis. Photosynthesis is the process by which higher plants manufacture dry matter through the aid of chlorophyll pigment, which uses solar energy to produce carbohydrates out of water and carbon dioxide. The overall efficiency of this critical process is somewhat low, and its mechanics are extremely complex. It is related to light intensity, wavelength, temperature, carbon dioxide concentration in the air, and the respiration rate of the plant. The distribution of solar energy within the plant community is affected by the leaf canopy's density, height, and capacity to transmit the energy; these therefore affect photosynthesis. The leaf-foliage density is characterized by the leaf-area index, the total leaf area of a plant over a given area of land. The optimum leaf-area index will vary between summer and winter and between temperate and tropical regions, but it represents a key factor in the search for better crop management based on improved photosynthesis. The efficiency of radiation utilization by field crops has been measured, showing that an ordinary crop converts less than 1 percent of available solar energy into organic matter.

Photoperiodism. Photoperiodism is another attribute of plants that may be changed or manipulated in the microclimate. The length of a day is a photoperiod, and the responses of the plant development to a photoperiod are called photoperiodism. Response to the photoperiod is different for different plants; long-day plants flower only under day lengths longer than 14 hours; in short-day plants, flowering is induced by photoperiods of less than 10 hours; day-neutral plants form buds under any period of illumination. There are exceptions and variations in photoperiodic response; also, it is argued that the truly critical factor is actually the amount of exposure to darkness rather than to daylight. Temperature is intimately related to photoperiodism, tending to modify reactions to daylength. Photoperiodism is one determining factor in natural distribution of plants throughout the world.

The phenomenon has many practical applications. Selection of a plant or a variety for a given locality requires knowledge of its interaction with the photoclimate. Artificial illumination is used to control flowering seasons and to increase production of greenhouse crops. In plant breeding, such stimulation of flowering has greatly reduced the time span from germination to maturity, shortening the time necessary to develop new varieties. In sowing field crops, photoperiodism can be used to select the date of sowing to produce optimum harvest size. Crop yield is reduced both by planting in a season that will cause plants to flower early and by planting at a time that will cause very late flowering. In Sri Lanka (formerly Ceylon), certain rice varieties with a vegetative period of five to six months may extend their life to more than a year when planted in the wrong season, causing almost complete loss of yield. Cowpeas in Nigeria will flower early and produce many seeds only when planted in daylengths of 12 hours or less.

WEATHER CONDITIONS AND CONTROLS

Three cardinal temperatures

Temperature. Regardless of how favourable light and moisture conditions may be, plant growth ceases when the air and leaf temperature drops below a certain minimum or exceeds a certain maximum value. Between these limits, there is an optimum temperature at which growth proceeds with greatest rapidity. These three temperature points are the cardinal temperatures for a given plant; the cardinal temperatures are known for most plant species, at least approximately. Cool-season crops (oats, rye, wheat, and barley) have low cardinal temperatures: minimum 32° to 41° F (0° to 5° C), optimum 77° to 88° F (25° to 31° C), and maximum 88° to 99° F (31° to 37° C). For hot-season crops, such as melons and sorghum, the span of cardinal temperatures is much higher. The cardinal temperatures may vary with stage of development. For example, cold treatment near 32° F (0° C) of germinated seeds before sowing can transform winter rye into the spring type; such treatment, called vernalization, has practical application in cold-climate plants.

The range of diurnal temperature variation is also important; the best net photosynthesis is related to a large diurnal temperature range, or high daytime and low nighttime temperatures. Knowledge of the difference between leaf and air temperatures aids farmers in adopting protective measures. In middle and high latitudes, frost often occurs before the air temperature drops to freezing; in summer, heat injury to plants might be much more serious than that suggested by the air temperature alone. Because of this factor, farmers in Taiwan shade the pineapple fruit to prevent heat damage.

Soil temperature sometimes is of greater ecological significance to plant life than air temperature. Germination of seed, root function, rate of plant growth, and occurrence and severity of plant diseases all are affected by soil temperature. Since an unfavourable soil temperature during the growing season can retard or ruin a crop, techniques have been developed for modifying the temperature. The two most important methods are (1) regulation of the energy exchange and (2) altering the thermal properties of the ground. Incoming energy can be regulated by an insulation layer on or near the ground surface, such as paper, straw, plastic, or trees; the outgoing radiation can be reduced by insulation materials or by generating smoke or fog in the air. Thermal properties of the ground can be modified by cultivation or irrigation, increasing the soil's ability to absorb radiation, or by varying the rate of evaporation. Mulching is a common technique for soil temperature control. Carbon black or white material can change the soil's ability to absorb radiation. In the Soviet Union, for example, it was reported that 100 pounds of coal dust per acre (112 kilograms per hectare) caused a one-month advance in the maturity date of cotton.

Frost. Another aspect of temperature control is frost protection. Likelihood of damage from freezing temperature depends upon the plant species, the season, the manner of temperature change, the physiological state of the plant, and other factors. Orchards can be located so as to minimize the chances of frost damage.

Radiation and wind types of frost

Two types of frost are recognized: (1) radiation frost, which occurs on clear nights with little or no wind when the outgoing radiation is excessive and the air temperature is not necessarily at the freezing point, and (2) wind, or advection, frost, which occurs at any time, day or night, regardless of cloud cover, when wind moves air in from cold regions. Both types may occur simultaneously. Most frost-protection techniques can raise the temperature only a few degrees, while some are effective only against radiation frost.

Heating is probably the best known and most effective frost-protection measure. It is most effective on nights with a strong temperature inversion, a condition in which the air temperature increases markedly from the ground up to as high as 40 or 50 feet (12 or 15 metres). The depth of air to be heated is thus rather shallow, and the area over which a given temperature rise can be produced increases linearly with the strength of the inversion. Lacking a temperature inversion, heaters protect by radiating heat to the plants and the ground surface, and by emitting a layer of humid smoke that reduces the net outgoing loss from the ground.

In general, a large number of small heaters is most effective; large heaters set up convection currents that break up the warm ceiling and draw in cold air. For radiation-frost protection, the heaters are placed in "view" of the plants or trees, but for advective frost the heavier concentration is placed along the upwind border. Common fuels for the heaters include oil, coal, briquettes, and wood. Oil is most effective, because it can be ignited rapidly and extinguished easily. Heating is a costly technique; a few growers who tried it in England soon gave up the practice, and, even in places such as California, heating is becoming less common and is mostly restricted to a few high-value crops such as citrus fruits.

Wind machines for frost protection

The wind machine is popular for frost protection; although it affords less reliable results, its operating cost is much lower than that for heaters. These machines, which are like fans or propellers, break up the nocturnal temperature inversion by mechanically mixing the air, returning heat to the ground that was lifted during the day. The stronger the temperature inversion, the more effective is the wind machine, which is ineffective, however, against a daytime freeze or cold soil. Even under the best circumstances, ground-surface temperatures will rise very little; therefore, some operators install both heaters and wind machines, using the latter for strong-inversion nights and the former for wind-frost protection.

Flooding and sprinkling with water prevent excessive ground cooling by increasing the heat conductivity and heat capacity of the soil and releasing latent heat of fusion, or the heat given off when the water freezes. The temperature of the plant will not fall below the freezing point so long as the change of state from water to ice is taking place. Flooding has the disadvantage of retarding increase in soil warmth during the day; thus, it can be used effectively for only one or two nights. Sprinkling creates water particles in the air that reduce outgoing radiation, but plant temperature declines immediately on cessation of sprinkling, and the ice formation may cause damage to the crop. In general, successful protection by flooding and sprinkling demands much skill and judgment from the operator.

Brushing is a frost-protection technique in which shields of paper or aluminum foil are set up to reduce radiation loss to the sky; it has been used with fair success for tomato culture in California.

Massachusetts cranberry growers add a thin layer of sand to the soil periodically. The sandy surface warms up easily and cools slowly by radiation; it also reduces evaporation of its low water content. Sanding can raise the temperature of loam, clay, and organic soils, thus diminishing frost hazard. Windbreaks can also function as frost protection by reducing inflow of cold air and by shielding plants from the total night sky.

Spraying of harmless foams or gels on plants threatened by frost is a technique under investigation. The trapped air in the foam serves as insulative protection, while the foam can be designed to dissolve after any desired time interval. The technique has been explored for use on strawberries and other low-growing crops.

Irrigation. Irrigation is probably the most common form of agricultural microclimatic control practiced by man. The basic techniques of irrigation are discussed in the section *Irrigation and drainage* above; at this point it only remains to make mention of efforts to correct deficiencies in precipitation, the deficiencies that lead farmers to irrigate.

Cloud modification versus rainfall

Rainmaking. Attempts to increase the amount of precipitation from clouds by seeding them with salt or silver iodide have been made for nearly three decades. Both aircraft and ground generators have been employed, but the techniques are typically beyond the means of an individual farmer. Results suggest that cloud modification is entirely possible, but the proof of increased rainfall at a level of statistical significance is a difficult problem. Success has been greatest under atmospheric conditions where natural rainfall is most probable. The prospect of modifying winter clouds to increase snowfall in mountain areas appears to be somewhat more promising, however.

Most cloud-seeding efforts are expended in regions where precipitation is only marginal for agriculture. It is commonly assumed that at least 20 inches (500 millimetres) of rain per year, fairly well distributed, is required to maintain a stable farming community. Unfortunately, the years of large deficiencies in such areas are those with only limited opportunity for cloud seeding. Some observers believe that weather modification to increase precipitation may yet become practical and economically feasible; the legal, ethical, and ecological problems raised by the prospect will not be easily solved, however.

Humidity. The value of high humidity in the greenhouse is well known, but knowledge of humidity–plant interaction under field conditions is comparatively slight. Other things being equal, the evapotranspiration rate decreases with increasing humidity; thus, rate of water use is higher at low levels of humidity. The benefits of irrigation are apparently greater when the humidity is high, which simply means that the efficiency of water use increases with humidity.

Wind. Wind affects plant growth in at least three significant ways: transpiration, carbon dioxide intake, and mechanical breakage. Transpiration (the loss of water mainly through the stomata of leaves) increases with wind speed, but the effect varies greatly among plant species; also, the effect is related to temperature and humidity of the air. In arid climates, dry and hot winds often cause rapid, harmful wilting. In winter, with frozen soil, the damaging effect of increased transpiration resulting from wind can be serious because the lost water cannot be readily replaced. By contrast, increasing wind promotes carbon dioxide intake within limits; this benefits the rate of photosynthesis. The effects of mechanical wind damage vary from species to species; some show a definite decrease in dry matter production with increasing wind, while others (usually short plants) are unaffected. Because of the long-recognized need, shelterbelts, massive plantings of trees that change the energy and moisture balance of the crop, are positioned to protect crops and to increase yields. A shelterbelt perpendicular to the prevailing wind reduces velocity on both sides. A medium-thick shelterbelt can reduce wind velocity by more than 10 percent to a distance of 20 times the tree height on the leeward side and three times the tree height to the windward. The length of the shelterbelt should be at least equal to that of the field to be protected. The sheltered area will suffer much less soil erosion and mechanical damage than unprotected areas. Other microclimatic effects of shelterbelts include: (1) small daytime temperature increases and nighttime decreases; (2) the occurrence of radiation frost in the leeside may be promoted; (3) rate of evaporation in the sheltered area is decreased, depending on wind velocity; (4) snow accumulates near the shelterbelt, causing increased moisture storage in dry farming.

Shelterbelts for crop protection

The overall effect of a shelterbelt is complicated but probably beneficial. There is much evidence that they increase efficiency of water use not only in subhumid and semi-arid regions but also in true deserts where oasis-type irrigation is practiced. The response to shelterbelts, however, depends on the species. Crops of low response to wind protection are the drought-hardy small grains and maize grown under dry farming conditions. Rice and forage crops such as alfalfa, lupine, and clover are moderately responsive. Crops that benefit most from wind protection are garden crops, such as lentils, potatoes, tomatoes, cucumbers, beets, strawberries, watermelons, deciduous and citrus fruits, and other tender crops, such as tobacco and tea. Some authorities assert that in strong wind areas shelterbelts will produce an average 20 percent yield increase, which is net gain of 15 percent when allowance is made for the land occupied by the belts themselves. Trees can be grown almost anywhere, even in the desert; tall plants such as corn (maize), sorghums, or even elephant grass can also be employed in arid regions by including them in the irrigation schedule. It would appear that windbreaks are among the most practical means of beneficial weather modification in agriculture.

The effects of pollution

Practically all forms of technology exact a certain price in environmental damage; agriculture is no exception. Agriculture in turn is sometimes damaged by undesirable by-products of other technologies (see also CONSERVATION OF NATURAL RESOURCES: *Pollution control*).

Air has physical properties and a chemical composition that are vital parameters of life for both plants and animals. Temperature, water vapour, movement, oxygen, and carbon dioxide in the atmosphere have a direct effect on food and fibre production. Air quality is changed by introduction of contaminants into it, and agricultural activities using such air may be affected adversely. Damage to plants by air pollutants is related to meteorological conditions, particularly temperature inversions in the atmosphere.

AIR POLLUTION

Air pollution damage to agriculture. For more than a century air pollution has affected agriculture. Burning coal and petroleum produce sulfur oxides. Fluorides result from smelting and glass and ceramic manufacture. Rising levels of ammonia, chlorine, ethylene, mercaptans, carbon monoxide, and nitrogen oxides are found in the air. Motor vehicles and growing population produce photochemical air pollution affecting not only the urban concentrations but also the contiguous rural areas. The mixture of pollutants from all sources, including agriculture, has released a host of contaminants into the air, such as aldehydes, hydrocarbons, organic acids, ozone, peroxyacetyl nitrates, pesticides, and radionuclides. The effect of these pollutants on food, fibre, forage, and forest crops is variable, depending on concentration, geography, and weather conditions. Damage to crops by air pollution, of course, brings economic loss as well.

Types of air contaminants

The effects of air pollution on plants and animals may be measured by the following factors: (1) interference with enzyme systems; (2) change in cellular chemical constituents and physical structure; (3) retardation of growth and reduced production because of metabolic changes; (4) acute, immediate tissue degeneration. Pollutants that enter the air from sources other than agriculture and that produce plant response are classified as: (1) acid gases; (2) products of combustion; (3) products of reactions in the air; and (4) miscellaneous effluents.

Acid gases. Acid gases include fluorides, sulfur dioxide, and chlorine. Hydrogen fluoride is extremely toxic to plants; some plants are injured by contact with concentrations of less than one part per billion. The damage apparently occurs initially to the chlorophyll, producing a mottled chlorosis and later killing the cells. Plants vary in degree of tolerance to hydrogen fluoride; usually the plants that accumulate fluoride readily are the most tolerant. Corn is more susceptible than tomato. All plants are most susceptible to fluoride injury during periods of rapid growth.

Sulfur dioxide given off in combustion of oil and coal commonly causes necrosis (cell death) of the leaf (Figure 10). At certain concentrations, sulfur dioxide will affect plants if the stomata (minute pores in the epidermis of a leaf or stem) are open. High light intensity, favourable growth temperatures, high relative humidity, and adequate water supply are conducive to open stomata. Plants that close their stomata at night can tolerate sulfur dioxide much better during that period. Conifers are more susceptible in spring and early summer, when the new needles are elongating. The sulfur dioxide absorbed by the leaf cells unites with water to form a toxic sulfite, but this is slowly oxidized to a relatively harmless sulfate. The toxicity of sulfur dioxide thus is a function of the rate at which it is absorbed by the individual plant; rapid absorption will cause greater injury. Chlorine damage to plants is somewhat rare; its typical symptoms are bleaching and necrosis of the leaf.

By courtesy of the U.S. Department of Agriculture

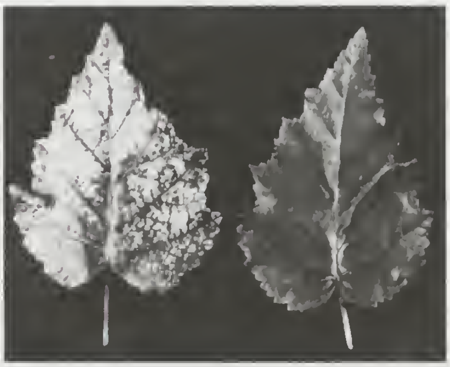


Figure 10: *Air pollution injury to vegetation.*
(Left) The effect of sulfur dioxide on a white birch leaf.
(Right) A normal leaf.

Products of combustion. The primary products of combustion are ethylene, acetylene, propylene, and carbon monoxide. Of these, ethylene is known to affect plants adversely; while the others may also do so, it would require higher concentrations of them than typically occur

Effects of ethylene

in polluted air. For many years it was observed that illuminating gas (3 percent ethylene) leaking from pipelines caused damage to nearby vegetation. Now, with the use of natural gas, ethylene in the air is derived mostly from certain chemical industries and from automobile exhaust. Greenhouse flowers in metropolitan areas are typically damaged by ethylene. Such injury appears to be caused by excessive speeding up of the life process, thus bringing on damage. Ethylene was first identified as affecting plant life over large areas in the field by its effects on cotton and other plants near a polyethylene factory.

Ethylene, ozone, and peroxyacetyl nitrate are produced as reaction products in the air and are clearly implicated in plant injury. In addition, certain bisulfites and nitrogen dioxide are under suspicion; there are probably others. Ozone is a major air pollutant affecting agriculture. Damage has been identified in a number of field crops, including spinach, tobacco, fruits, vegetables, forest trees, and ornamentals. Symptoms of ozone toxicity appear as flecks, stipple, streaks, spots, tipburn, and premature yellowing of the foliage; these may be visible only on the upper leaf surface. Peroxyacetyl nitrate and its analogs produce symptoms called silver leaf and leaf banding, which have been observed in the Los Angeles area and elsewhere for many years.

The adverse effects of airborne radioactive contaminants on the agricultural economy at the present time are small.

Air pollution by agriculture. Contributions of agricultural technology to air pollution include pesticides, odours, smoke, dust, allergenic pollens, and trash. The widespread public concern about pesticides makes it imperative that pesticide technology be carefully controlled and that search for better methods be pursued vigorously.

Persistence in pesticides

Pesticides. The problem of persistence in pesticides can be highlighted by noting that this attribute exists in a range from moderately persistent (a lifetime of one to 18 months—2,4-D, atrazine); persistent (lifetime up to 20 years—DDT, aldrin, dieldrin, endrin, heptachlor, toxaphene); or permanent (lead, mercury, and arsenic). Presumably, the less persistent types should be more desirable, other things being equal; but those that degrade rapidly, such as the organophosphate insecticides, are extremely toxic and nonselective, which encourages rapid emergence of resistant insects and destroys their natural enemies. Thus, it is apparently not possible to adopt chemicals that function without some drawback or disadvantage.

Whether pesticides are applied by spraying or by surface application, air is the usual medium through which the chemicals move to their intended and unintended targets. Reliable data on how pesticides behave in air, such as distance travelled, are lacking, because adequate monitoring is unavailable. Their chemical and physical nature, method of application, and the atmospheric conditions will influence their concentration and ultimate fate. There is no doubt that pesticides may be transported long distances on dust particles. The rate of removal from air is difficult to predict, but in the long run the chemicals return to earth.

Odours, pollen, and dust. Odours from animal concentrations are recognized as being highly undesirable to air quality. Where these operations exist contiguous to urbanized areas, public reaction is usually unfavourable. Disposal of animal waste on the land may worsen the odour problem; in addition, high wind may move dry increments into the air. Smoke is emitted by operations designed to dispose of crop residues, or by controlled burning of weeds and brush. Air quality is also affected by transmission of allergenic pollen such as ragweed pollen, which can be blown for hundreds of miles.

Improper land use and treatment can cause considerable deterioration in air quality. Practices that strip the soil of plant growth or crop residues for long periods contribute to wind erosion, particularly in dry-farming areas. Fortunately, the technology of preventing wind erosion is well understood and widely used. Trash related to agriculture is moved freely by wind and distributed in unwanted fashion. Hulls of rice and wheat and cotton-gin trash are examples of this kind of airborne nuisance.

In contrast to most other technologies, however, the agricultural variety offers a major beneficial contribution to air quality. The photosynthesis of green plants removes carbon dioxide from the air and adds oxygen to it, thus helping to maintain the life-giving balance between these gases.

SOIL AND WATER POLLUTION

Pollutants damaging to agriculture. Soil and water pollutants that may adversely affect agricultural operations include sediment, plant nutrients, inorganic salts and minerals, organic wastes, infectious agents, industrial and agricultural chemicals, and heat.

Sediments in dual pollution role

Sediment. Sediment is a resource out of place whose dual effect is to deplete the land from which it came and impair the quality of the water it enters. Aside from filling stream channels, irrigation canals, farm ponds, and irrigation reservoirs, sedimentation increases cost of water clarification. Suspended sediment impairs the dissolved-oxygen balance in water. The recreational value of farm ponds is diminished by sediment, while soil depleted farmland is reduced in value.

Plant nutrients. Nutrients of plants become resources out of place when they appear in groundwater and surface water; in fact, they become serious pollutants. Unwanted aquatic plants are nourished by plant nutrients derived from agricultural runoff, feedlots and barnyards, municipal and rural sewage, and industrial wastes. Aquatic plants clog irrigation and drainage structures, thus increasing maintenance cost and reducing capacity. Nitrates and nitrites in groundwater, which can poison human beings and livestock, result from both agricultural and industrial operations.

Inorganic salts and minerals. Inorganic salts and minerals that impair the quality of soil and water are derived from natural deposits, acid mine drainage, industrial processes, and drainage flow from irrigated areas. Salt accumulation on irrigated soils causes the most damage and loss in this category. A high proportion of sodium in irrigation water supply affects plant life adversely (see below *Salinity*). More than just a trace of boron is highly toxic; therefore, water used in municipal and industrial processes involving borax may not be usable for agriculture.

Organic wastes. Organic wastes emanating from municipal sewage, garbage, food-processing industries, pulp mills, and animal enterprises are attacked by aerobic bacteria. When this occurs in water, the oxygen content of the water is depleted or reduced to zero, at which point the anaerobic bacteria complete the process of reducing the wastes to inert material. This produces septic conditions that make the water unfit for recreational use, farmstead supply, or crop irrigation.

Diseases transmitted by water and soil

Infectious agents. Where not carried by wind, infectious agents are transmitted mainly by water and soil. Bacterial and virus diseases of crops are spread by machines that move contaminated soil. Insects are prime carriers of these diseases. Weed seeds are spread by irrigation water, as are nematodes. Animal diseases transmitted by water and soil include leptospirosis, salmonellosis, hog cholera, mastitis, foot and mouth disease, tuberculosis, brucellosis, histoplasmosis, Newcastle disease, anthrax, coccidiosis, and many others. Mosquitoes breeding in stagnant water can transmit encephalitis. Most crops and livestock in the world are susceptible to one or more highly infectious disease that may be transported by soil or water. The cost of losses from these diseases is staggering.

Chemicals. Organic chemicals in soil and water, such as detergents, insecticides, herbicides, fungicides, nematocides, rodenticides, growth regulators, and defoliants, can have adverse effects on agriculture. The application of persistent insecticides to potato lands has led to residues in sugar beets grown in the same soil the following year, for which there are no tolerances. Fish have been killed in farm ponds because of drainage of insecticide pollutants. Use of heptachlor (no longer recommended) to control alfalfa weevil led to soil contamination and uptake by hay; dairy cows that ate the hay produced milk containing heptachlor.

Aerial and ground application of herbicides on nonagricultural lands (utility rights-of-way, roadsides, industrial sites) often cause damage to nontarget crops. Herbicide wastes may enter drainage or irrigation ditches and create trouble. The presence of chemical residues in agricultural commodities can cause serious problems ranging from confiscation to loss of public confidence. Practically all aspects of chemical usage are now regulated or restricted by government.

Heat. Introduced into water by industrial processes, heat can have a detrimental effect on fish and other creatures in the water; damage to recreational value can result. But, though heat is a water pollutant, its effect is minor with respect to agriculture.

Pollutants from agriculture. Some pollutants from agriculture have adverse effects on agriculture itself, as excesses of plant nutrients and salts from irrigation. These pollutants and others also affect the environment at large.

Eutrophication. Eutrophication occurs in a body of water when an increase of mineral and organic nutrients has reduced the dissolved oxygen, producing an environment that favours plant over animal life. The resulting algae and other water plants tend to choke other forms of life in the oxygen competition, especially where carbon and phosphorus are plentiful. Doubtless, much phosphorus in streams and lakes is delivered from agriculture, but primarily through soil erosion rather than runoff. Though the principal source of phosphorus is apparently municipal sewage-treatment plants, direct runoff from feedlots may also contain large amounts. The solution to the problem of phosphorus in surface water lies in using good soil-conservation practices and in minimizing runoff from animal concentrations and manure.

In contrast, identification of nitrate sources in water supplies has suffered from conflicting evidence. Where nitrate is found in water, some have concluded it came from chemical fertilizers, while others have suggested it came from natural soil nitrification or nitrification of sewage effluent or animal wastes. The problem has serious aspects, because nitrate can cause serious illness in human beings. One difficulty in identifying nitrogen sources lies in the fact that it is present in soils for reasons other than fertilization; the growth of legumes, for example.

Salinity. Salinity is a major problem in irrigation agriculture. Through evapotranspiration, salts in the irrigation water become more concentrated in the drainage effluent. It is therefore claimed that water quality is seriously impaired by irrigation agriculture. Irrigation water always contains some salt, most of which is excluded by the plant roots; since the evaporated water is pure, the soil accumulates the residual salt, which is added to what the arid soils already have in abundance. This accumulation of salt must be removed if plants are to be grown at all, and it is removed by leaching with excess water.

Survival of an irrigated arca will depend, therefore, on a favourable salt balance: salt leaving the area must equal or exceed that received in the water supply. The irrigation farmer is not actually "producing" a contaminant but is transferring one in a more concentrated form. Future intensified use of limited irrigation water may add to the severity of this problem. Where the return flow is readily recoverable, as from tile drains or pumped wells, it could be purified by a desalination process, returning the purified fraction to the watercourses and disposing of the concentrated-salt fraction in such a way that usable groundwater is not affected.

Agricultural processing wastes. The wastes from processing of agricultural products represent another pollution hazard. These include runoff or effluent from sawmilling, pulp manufacture, fruit and vegetable canning, cleaning of dairies, slaughtering of meat animals, tanning, manufacturing of cornstarch and soy protein, sugar refining, distilling, wool processing, and many others. The runoff from agricultural enterprises can contain disease organisms and other infectious agents. Insects associated with agriculture can transmit diseases. Plant diseases move from agriculture to lawns, gardens, parks, and golf courses.

Monitoring pesticides. The monitoring of pesticides in water has been carried on in various areas since World War II. Some of the monitor networks, backed by analysis laboratories, are quite extensive. The accumulated data show how and when certain pesticides move from target areas into other parts of the environment. Ponds and catch basins sometimes show measurable amounts of pesticide rcsidues from water leaving fields. Although most organic insecticides are hydrophobic and almost insoluble in water, they can become attached to materials suspended in water; but, after these materials settle, the rcmaining amounts of insecticide residues usually becomc ncgligible. This confirms the earlier supposition that the movement of chemicals from target areas is greatest when silt and organic loads are high in runoff watcr.

Transport pattern of pesticides

Levels of pesticides in soils arc constantly changing. So many variables and processes are involved, that rates of accumulation of even the most persistent insecticides are quite variable and difficult to detcrmine. Soil monitoring programs are underway, howcvcr, and are providing much-needed information. The problem of accumulation in soils arises because the tiny organisms in soil are not capablc of degrading many pesticides at rates sufficiently high to prevcnt soil and also water pollution. Thus, the persistent types, such as DDT and other chlorinated hydrocarbons, remain available for absorption by higher animals (including human beings) and for causing harm to nontarget organisms. (R.E.S.)

PRINCIPLES AND PRACTICES OF ANIMAL HUSBANDRY

Animal breeding

Animal breeding is the controlled propagation of domestic animals. Its aim is the improvement of qualities considered desirable by man. Breeding procedures involve the application of several basic sciences, chiefly reproductive physiology, genetics, and statistics. This article deals with the practical application of scientific principles to the selection of superior animals and the planning of mating combinations. The fundamental biological principles underlying animal breeding are discussed in the articles GENETICS AND HEREDITY and REPRODUCTION AND REPRODUCTIVE SYSTEMS.

Uses of highly bred animals

Animals are bred for utility, sport, pleasure, and research. Dogs, for example, have been bred to serve as watchdogs, police dogs, hunters, sheep dogs, and pets. Many species of small mammals, especially rats, mice, rabbits, and guinea pigs, are bred for research, chiefly in genetics, physiology, and medicine. The basic principles of breeding are the same no matter what the animal species or the purpose in breeding may be, but the practical approach to the problem may differ in several respects, depending on such considerations as the mode and rate of reproduction and the relative effects of genetic and cnvironmental factors on the traits of greatest interest.

The term population is used in this article to denote a group of interbreeding individuals; *i.e.*, a breed, or strain within a breed, which in some respects is genetically different from other breeds, or strains, of the same species. The word purebred is used here in its scnse of referring to animals registered in the herdbook maintained for a certain breed, or to animals eligible for such registration, and the mating of purebred animals is called pure breeding. It is to be understood that genetically pure breeds (homozygous for all traits) do not exist.

The objectives of animal breeding vary with regard to species, local conditions, and time. Early in history horses were bred mainly for riding or loading purposes; later they were bred for traction; and nowadays, to a large extent, for sport (racing and hunting). In North America and western Europe, cattle populations are specialized for beef or milk production, or bred for a combination of both. In southern Europe and in many parts of Africa and Asia, oxen are still produced for pulling plows or carts. Some breeds of sheep are specialized for wool production, some for meat, and one breed, the Karakul sheep, is bred for fur production (Persian lamb). Pigs always are bred for meat production, but they may be specialized to produce a certain type of meat, either pork or bacon. At one time, chickens were bred for the combined production of eggs and meat, but in the Western world there is now a pronounced specialization of breeds and crosses to produce either eggs or meat.

EVALUATION OF ANIMALS

In breeding farm animals for utility, selection must be based as far as possible on objective measurements of traits that are decisive for the economy of production. Judging animals on the basis of appearance alone has not become obsolete but its importance has been reduced. Certain traits, of course, are difficult or impossible to measure objectively. In all kinds of horses, for example, the legs should be well proportioned in relation to the body and free from faults and weaknesses; this judgment is difficult to render quantitatively. Similarly, the strength of the legs is important in cattle and pigs, but difficult to measure. In dairy cows, the attachment of the udder to the body is difficult to measure; but it is significant, and it can be judged and scored. In mink breeding, a judging of the animals with regard to fur quality and shade of colour is necessary and cannot, so far, be substituted by laboratory tests. In the breeding of pet animals, such as dogs and cats, judging the animals for conformity with breed standards is usually decisive in determining their market value.

Judging animals on appearance

Actual measurement of an individual animal's performance is a fairly recent innovation in animal breeding, except with regard to Thoroughbred horses, the selection of which for centuries has been based on speed at standard racing distances. Systematic recording of milk production in dairy cows started in Denmark in 1895, and the movement spread rapidly in northwestern Europe and North America. In several countries, young beef and dairy bulls are tested at special stations for rate of growth and muscle development. In sheep breeding, the number of lambs borne by each ewe is recorded, as is the weight of the lambs at five months of age. The recording of growth rate, feed consumption, and carcass quality of pigs began in 1907. Progeny testing and performance testing (see below *Methods for estimating breeding values*) has subsequently been added to judge young males (bulls, rams, and boars) intended for breeding. The productivity of sows is measured by recording the number of pigs in each litter at birth and their weight at three weeks of age. An ultrasonic technique measures the thickness of backfat on live pigs. Special nests with hinged doors (trapnests) were employed to record the egg production of individual hens from the end of the 19th century until the introduction of individual laying cages made trapnests unnecessary. The recording of individual performance in breeding populations of farm animals developed with great rapidity after World War II, in terms of both frequency of use and refinement of methods. It may be considered to be the very foundation of progress in breeding programs.

Judging animals on performance

Genetic and environmental variation in animal traits. The traits of animals in any population may be classified roughly into two groups; viz., qualitative and quantitative traits. Qualitative traits show discontinuous variation; for example, coat colour, presence or absence of horns, certain blood characteristics (*e.g.*, blood types), presence or absence of particular enzymes, and existence of metabolic defects and congenital malformations (*e.g.*, bleeding disease [hemophilia] in dogs). In general, the inheritance of qualitative traits is relatively simple, in accordance with the laws of heredity, and environment plays only a minor role in their variation.

The quantitative traits show continuous variation between the extreme variants, the mean type being, as a rule, most frequent. Growth rate, live weight, body measurements at mature age, milk yield and composition in cows; body length and backfat thickness in pigs; wool yield and quality in sheep; and egg production in fowl are typical examples of quantitative traits. As a rule, such traits are influenced by many genes (polygenes), each gene exerting relatively small effect, and environmental factors are responsible for a considerable part of the variation.

There is, in fact, no distinct borderline between qualitative and quantitative traits. A congenital malformation, for example, may be determined by a major gene, as well as a number of polygenes, and the latter may cause a considerable variation in the expression of the trait. Similarly, variation in the expression of a trait regulated by a single gene may be caused by environmental factors. Nevertheless, this classification is useful because typical qualitative traits can be analyzed with regard to a single gene and its behaviour; whereas quantitative traits are best studied by statistical methods that permit one to proceed without knowledge of the number of genes involved or their interrelations. By using appropriate statistical methods, for example, it is possible to estimate the fraction of the total variation in the population that is caused by the additive effect of the genes; this fraction is termed the heritability. A heritability of 1.00 indicates that all the variation observed in the population of the trait in question is genetically determined; a heritability of 0.00 indicates that the variation is wholly environmental in cause; and when the heritability is 0.50 the genetic and environmental factors are equally responsible for the observed variation of the trait.

A few estimates of heritability of traits are given to illustrate typical values. In horses: pulling power, 0.25. In dairy cows: height at withers at three years of age, 0.45; milk yield at first lactation, 0.30; fat and protein content of milk, 0.55; resistance to mastitis, 0.30. In beef cattle; daily weight gain (from weaning to slaughter), 0.40. In sheep (Merinos): yield of clean wool, 0.45. In pigs: daily weight gain (from weaning to slaughter) in group feeding, 0.30, in individual feeding, 0.60; backfat thickness, 0.50; resistance to atrophic rhinitis, 0.25. In chickens: egg production (in first laying year), 0.30; egg size, 0.50; resistance to leukosis, 0.10. In general, heritability is relatively low for such traits as fertility and resistance to infectious diseases, and it is high for growth rate, body size at mature age, and composition of cows' milk.

Heritability of traits

The magnitude of environmental effects, and to some extent also the additive genetic effects, may vary considerably between different populations of the same species; therefore, heritability estimates also vary and cannot be considered as constants for the various traits. For example, as noted above, when the heritability of growth rate of pigs is estimated from individual-feeding data, it is twice as high as when it is estimated from group-feeding data. In spite of these limitations, heritability estimates for quantitative traits have proved to be of great value in planning animal improvement programs.

Methods for estimating breeding values. The breeding value of an animal depends on the genes it passes on to its offspring. Each individual offspring receives a random sample of one-half of the genes from each of the parents.

Concerning qualitative traits with known inheritance patterns, the situation is relatively simple. A dominant gene manifests its presence in the physical character (phenotype) of the individual even when the gene it is paired with is different—that is, in the heterozygous condition. A recessive gene also can be recognized but only when it is present in duplicate (the homozygous state). In both cases, of course, there must be no complications caused by modifying polygenes or environmental factors. The breeder can then select for or against the trait according to his wishes. To recognize heterozygous carriers of recessive genes by visual inspection of the animals is impossible in most cases, but it can be done by mating the suspected carrier to a certain number of known heterozygotes or to its own offspring.

The breeding value of a quantitative trait may be assessed on the basis of the merits of: (1) the individual himself, (2) his ancestors, (3) his collateral relatives, (4) his progeny, or (5) a combination of any two or more of the above. The relative value of these different approaches depends on the heritability of the trait and the rate of reproduction. With regard to quantitative traits that are sensitive to environmental influence, it is often advisable to use the deviation of the individual's record from the mean record of the herd, rather than the individual's record itself. This procedure eliminates, or at least reduces, the effect of

Breeding values

nongenetic differences between herds. Each of the above methods of assessing quantitative traits has its own uses (as discussed below).

Individual merit. If it is assumed, for example, that a cow in her first lactation has produced 10,000 pounds (4,500 kilograms) of milk, and that her contemporary herd mates in the corresponding lactation have produced 8,900 pounds (4,000 kilograms) on average, then the phenotypic merit of this cow is 1,100 pounds (500 kilograms) above that of her contemporaries. By this procedure, environmental effect is minimized. When cows of different age (or undergoing different lactation periods) are compared, it is necessary to correct the yield of each individual to a standard age, because the yield increases until the fourth or fifth lactation. The fat content of the milk is much less influenced by environment and by individual age; therefore, no age corrections are needed, and individuals can be compared directly on their actual records. The same principles apply to other traits and other species of animals.

Ancestor merit. The merit of ancestors is usually the first available information on the breeding value of an individual, and such pedigree information, therefore, is valuable as a rule. With each earlier generation in the pedigree, however, the value of this information is halved. Furthermore, since a grandparent can pass his genes to a grandson or a granddaughter only through one of its parents, the more information known about this parent, the less valuable is the information known about the grandparent. For example, if a reliable progeny test (see below) shows that the sire of an animal has a high breeding value, there is no need to consider the parents of that sire. In many cases, especially in horse and dog breeding, the importance of long pedigrees has been greatly exaggerated.

Judging by sibs

Merit of collateral relatives. The genetic similarity between an individual and a randomly chosen full sib (brother or sister) is the same, on average, as that between the individual and one of his parents; and an individual's genetic similiarity with a half-sib is the same as that with one of his grandparents. An individual, however, can have many more full sibs and half-sibs than parents or grandparents, and the sibs, therefore, may be of much greater value than parents or grandparents for estimating the breeding value of the individual. In pigs, rabbits, dogs, and fowl, the number of full sibs can be fairly large, and in artificial insemination of cattle the number of half-sibs can be very large.

Progeny merit. Progeny tests yield the final information on an animal's breeding value. The relative importance of such tests increases with decreasing heritability of the trait, and it is especially valuable for sex-limited traits, as, for example, in testing a bull's breeding value with respect to the milk yield of his daughters. The lower the heritability of a given trait, the larger a progeny group is needed for a reliable test of the individual.

Combined methods. Information from two or more of these methods can be combined into a single estimate of the individual's breeding value. In such circumstances, the different criteria should be weighted according to their expected contribution to the accuracy of the final estimate.

THE BREEDING PROGRAM

The genetic improvement of a herd or a breed requires careful planning with regard to the choice of animals for breeding and the mating combinations that are carried out.

Methods of selection. Selection of breeding animals can be carried out in different ways. Among the more important are mass selection, pedigree selection, family selection, and progeny selection.

1. Mass selection is based solely on individual merit. Applied to traits with high heritability and about equal manifestation in both sexes, mass selection can be expected to give good results. With decreasing heritability the efficiency decreases, and for sex-limited traits (and heritability considerably below 0.5) it is always inefficient.

2. Pedigree selection depends on the merits of the ancestors. It is valuable in the first selection among young animals, especially when the heritability of the traits is high. Relying solely on pedigree selection, however, results in very slow progress.

Family selection methods

3. Family selection is based on the merits of collateral relatives, such as full sibs or half-sibs, and it is used mainly as an aid to individual selection. It is especially valuable with regard to sex-limited traits, for traits with low heritability, or when some animals have to be slaughtered, as for determining the carcass quality. In the selection of young males for breeding, for example, no data may be available on their individual performance; *e.g.*, egg production in the fowl. When sib groups of pullets start laying early in the autumn, the cockerels may be selected for breeding on the laying records of their full sibs and half-sibs. Similarly, young bulls may be selected mainly on the milk records of their paternal half-sibs; that is, on the progeny tests of their sires. Individual pigs may be selected on the basis of carcass tests made on their sib groups. With regard to traits that can be recorded for all the animals alike, males and females, selection of individuals can be based simply on their family average or on an index that combines the individual's own performance and the average for the rest of the family. A simple procedure, applicable in dog and pig breeding, would be to select the best individual from the best litter.

One difficulty in the application of family selection is that systematic environmental differences may occur, especially between full-sib groups, and these tend to mask the genetic differences. When inherited defects appear in sib groups, there is a certain risk that some of the healthy animals carry a hidden gene for the defect.

4. Progeny selection has been applied with great success in dairy cattle breeding, and in general it is valuable in all types of livestock when applied to sex-limited traits and traits with low heritability. Early progeny testing of males on a sufficient number of offspring and an effective selection among those tested are very important. The disadvantage of selection of sires on progeny testing is that it means increased length of the generation interval and thereby tends to slow down the rate of genetic improvement.

In general, it is not necessary to include all the traits used for selection in an overall index; for example, all animals used for breeding should possess normal fertility, and those that do not should be excluded from the breeding program. Also, as a rule, any animal known to be a carrier of a gene for a serious metabolic or morphological defect should be eliminated even if the merits for some other traits are fairly high.

Stepwise selection

Usually selection is made in a stepwise fashion. With regard to dairy bulls, for example, selection on the basis of pedigree can be made soon after birth; a second selection can be made later based on growth rate during the first year of life and fertility in the first series of inseminations; and finally a third selection can depend on the results of progeny testing, when offspring are old enough to be judged.

Mating systems. *Random mating.* Random mating implies that each possible mating in a population has the same probability of occurrence. Only artificial selection by the breeder can really be eliminated; however, a certain amount of natural selection always takes place. Random mating often is used in breeding experiments to minimize genetic changes in a control population with which selected populations are compared.

Inbreeding. Inbreeding may be defined as mating of individuals more closely related than the average of the population. It increases the homozygosity and decreases the heterozygosity of the inbred animals. The so-called inbreeding coefficient is a measure of the loss of heterozygosity due to inbreeding, and it is expressed as a fraction, or percentage, of the amount of heterozygosity present when inbreeding started. After one generation of mating between full sibs or mating of sire with daughter or dam with son, the heterozygosity of the offspring is reduced by 25 percent (or the inbreeding coefficient is 25 percent). In the mating of half-sibs or double first cousins the inbreeding coefficient is 12.5 percent. Mating of single first cousins gives an inbreeding coefficient of 6.25 percent, and that of half first cousins 3.12 percent. Mating between full sibs in two successive generations decreases the heterozygosity by 37.5 percent, and in three generations by

50 percent. So-called inbred lines are produced by continuous consanguineous matings in several generations; genetic variation decreases within each line and increases between separate lines. In experiments with mice, rats, and guinea pigs, full-sib matings have been continued through many generations. Farm animals and birds are much more sensitive to inbreeding, and usually full-sib matings can be continued for a few generations only because of a marked decrease in viability and fertility.

Inbreeding in small populations

Breeding within small populations, such as a herd or flock, without infusion of new animals from outside, leads automatically to a certain amount of inbreeding. In farm mammals, each male is invariably used to serve a large number of females, and under such conditions the increase of the inbreeding coefficient per generation in a closed breeding unit can be estimated simply by dividing 100 by eight times the number of males used. Thus, when only one homebred sire is used, the decrease in heterozygosity in each generation is 12.5 percent. When artificial insemination is used, a small number of sires again are used to serve a large number of females. Nevertheless, the reduction in heterozygosity may be relatively small in such cases, because many young sires are used for progeny testing, and several tested sires enter into service each year. The modern trend in artificial insemination is to use the tested sires for only relatively short periods of time, after which they are replaced with younger sires, which presumably have made even better records. Such rapid turnover in sires serves also to reduce the length of the generation interval in breeding programs.

Inbreeding increases the homozygosity of unfavourable as well as favourable genes. As a result there is a segregation of various kinds of congenital defects and, more important, a general decline in fertility and viability of the inbred animals. The latter finding has been demonstrated in numerous experiments with farm animals, especially pigs and poultry. In experiments with laboratory animals it has been shown also that the sensitivity to unfavourable environmental influences increases with inbreeding. Although the decrease in fitness resulting from inbreeding is a general phenomenon, the amount of decrease is dependent on the genetic constitution of the animals used for inbreeding.

Linebreeding is a form of mild inbreeding designed to concentrate the genes of a certain ancestor in a strain of animals. The most intensive form of linebreeding is repeated backcrossing to a particular parent, but usually a more distant relationship is preferred; for example, a female may be mated to her grandsire or uncle.

Outbreeding. Outbreeding is defined as mating individuals less closely related than the average of the population. The degree of outbreeding can vary, just as that of inbreeding can. In some cases, it is even possible to make crosses between species, as in the crossing of the horse and ass for the production of mules. The term crossbreeding, however, usually refers to crosses between breeds within the same species. Crosses often are made between more or less inbred strains or lines of the same breed (called, respectively, strain and line crosses).

Crossbreeding

Crossbreeding has been practiced for a long time, and it also has been subjected to experimental research in the United States and Europe. Various methods have been developed and tested, with generally favourable results. The main function of crossbreeding is to increase the heterozygosity of the offspring. One of the breeds selected may be superior in certain traits and the other breed may excel in other traits. It can be expected, then, that the first-generation crossbred animals will be about intermediate to the parental breeds with respect to both traits. In some cases, however, the first-generation animals are somewhat superior to the better parental breed with regard to total merit. When the average quality of the first generation exceeds the average of the two parental breeds, the phenomenon is called heterosis, or hybrid vigour. Heterosis is displayed mainly in the so-called fitness traits, fertility and viability. It is thought to result from interaction of different forms (alleles) of a given gene (a phenomenon called overdominance) or from interactions of quite separate genes (epistasis). Crossbreeding thus is a way of utilizing nonadditive gene effects that cannot be exploited by selection within the separate breeds. When the heterozygosity is increased, the number of different genes is increased, a result that probably makes the animals better able to adapt to environmental stress.

Rotational crossbreeding

Backcrosses (crosses of crossbred offspring to one of the parental breeds) and successive three-breed crosses seem to have an advantage when maternal influence on the offspring is important; crossbred sows, for example, generally take better care of their young than purebred sows do. Rotational crossbreeding, usually involving three breeds, is a favoured method in commercial pig breeding since it necessitates purchase of males only. The first cross is made between breeds A and B, for instance, and the female offspring are mated to a boar of breed C; the females resulting from this cross are then mated to a boar of breed A, and in the next generation a boar of breed B is used, and so on.

Many experiments have been carried out on the crossing of inbred lines of the same breed or of different breeds. Generally, fertility and viability are restored even in the first crossbred generation. It has been possible with this technique, for example, to produce commercial "hybrid chicks" with superior egg-laying performance. Because the individual inbred lines are poor producers, the hybrid chicks are usually developed by a four-way cross. This is carried out by mating the offspring from crossing of lines A and B with the offspring from crosses between lines C and D, producing in effect a "double" hybrid. In order to obtain the best possible result, a large number of lines are tested in various crosses for combining ability—that is, ability to produce desirable results from crossing. Most of the eggs marketed in North America and western Europe are produced by hybrid chicks.

Similar methods of breeding have been tried also on pigs and cattle, but the results have been less favourable. With these animals the market value of the individual is so high that it is impossible for breeders to keep the costs of developing and testing inbred lines within reasonable limits.

Preservation of semen

Artificial insemination and egg transplantation. The practical application of artificial insemination in horses, cattle, sheep, and pigs was developed in Russia during the first decades of the 20th century. Semen was collected in an artificial vagina when the male mounted a female, or a dummy, and methods were developed for dilution of semen and its preservation for several days outside the body. More rapid progress became possible in 1950, when it was shown that bull semen could be deep-frozen at −79° C with solid carbon dioxide and later thawed without serious effect on fertility, provided that a certain amount of glycerol was added before freezing. Later, the use of liquid nitrogen made it possible to store semen at about −196° C. Calves have been produced from semen that has been frozen for more than 10 years, and it seems possible to increase the period of storage indefinitely. Theoretically, it should be possible to produce more than 100,000 calves per bull annually; 10,000 or more actually have been produced. In Denmark about 95 percent of the dairy cows are artificially inseminated; in England and Wales, about 70 percent; and in the United States, about 50 percent.

The greatest advantage of artificial insemination of dairy cattle from a genetic point of view is that the bulls can be progeny tested with much greater accuracy than in natural breeding, not only because the number of daughters is greater but also because the daughters are spread over many herds with different environmental conditions. A progeny test carried out in any one herd is valid for the conditions in that herd only. Furthermore, by artificial insemination the progeny test can be completed at least one year earlier, on average, than if the bulls were used in natural service.

Artificial insemination is practiced also on beef cattle, horses, sheep, pigs, and poultry, although on a small scale compared to its use with dairy cattle. The application of artificial insemination in pig breeding was delayed by difficulties in deep-freezing boar semen, but these difficulties have been overcome.

Polyovulation

By hormone treatment, production of many ova (polyovulation) can be induced in females. After artificial insemination, the fertilized eggs can be collected from the oviducts or uterus of the donor and transferred to the uteri of recipient females for development into normal fetuses. At the time of transfer, the recipient must be synchronized with the donor with regard to the sexual cycle. Many successful transplantations of fertilized eggs have been made in sheep, and some have been carried out in cattle and pigs. Theoretically, it is possible to harvest thousands of eggs from an individual female within a few years' time and to distribute these eggs to a large number of recipients, thereby multiplying the possibilities for propagation of superior females. To secure the eggs from the donor, however, and transfer them to other females is laborious and costly, even if it can be done without major surgery. Methods of freezing and storing eggs without harmful effects must be developed before the transplantation technique can be applied as a routine procedure in the breeding of farm animals.

The rate of genetic improvement. The rate at which genetic improvement can be carried out within a certain population depends on three factors:

1. The accuracy in estimating the breeding value of the individuals. This accuracy itself depends on the heritability of the trait or traits on which the selection is based. The value used may be based on the animal's own performance, as, for example, the milk yield of a cow, or it may be an index formed by a combination of information on breeding value derived from various sources such as pedigree, collateral relatives, and progeny.

2. The selection differential, which is defined as the difference between the index of the animals selected for breeding and the mean of the entire population or group to which the selected animals belong.

3. The length of the generation interval, defined as the average interval (in years) between the birth of the parents and the birth of their offspring used for breeding. On average, this interval for horses is about nine years, for cattle five, for sheep three and a half, for pigs two and a half, and for chickens one and a half. It is possible, however, to reduce these intervals quite considerably by planned breeding.

The overall rate of genetic improvement, or response to selection, is equal to the accuracy in selection multiplied by the selection differential and divided by the generation interval. In equation form this is expressed as: $Re = HS/Y$, in which Re is the response to selection, H is the heritability, S the selection differential, and Y the generation interval.

Response to selection

From this formula, it is evident that the rate of genetic improvement of the population can be increased by increasing the accuracy in choice of breeding animals, by increasing the selection differential, and/or by decreasing the length of the generation interval. The possibility of increasing the selection differential depends on the rate of reproduction. Pigs have a much higher reproductive rate than cattle, and therefore fewer animals are needed for breeding and a greater selection differential is possible. In dairy cattle it may be necessary to raise about 60 percent of all heifer calves for breeding, merely to keep the size of the herd constant. For bulls, however, the selection differential can be very high, especially when artificial insemination is used. The size of the population subjected to selection is also important. It has been calculated, for example, that when the number of performance-tested dairy cows in an artificial insemination unit increases from 2,000 to 20,000, it should be possible to increase the rate of genetic improvement for milk yield about 50 percent by means of efficient progeny testing and selection.

In most Western countries there was a pronounced improvement in many economically important traits of farm animals after World War II. In the United States, for example, the average yield of milk from milk-recorded cows increased from 9,425 pounds (4,275 kilograms) in 1955 to 12,209 pounds (5,538 kilograms) in 1967, an increase of 29.5 percent, or 2.27 percent per year; by 1980 an official test reported an average yield of 14,960 pounds (6,786 kilograms). For the Swedish red and white breed, the average milk yield increased from 9,473 pounds (4,297 kilograms) in 1960 to 11,094 pounds (5,032 kilograms) in 1969; *i.e.*, by 17.1 percent, or 1.71 percent per year. It has been estimated that the genetic improvement of this breed during the same period corresponded to about 1 percent per year, or more than half the actual rise in milk yield. It seems probable that the major part of the genetic improvement in this instance was due to a more efficient progeny-testing and selection program that had been started among the tested bulls in the mid-1950s. Similarly, in Danish pig-testing stations, the average daily gain in live weight increased from 23.9 to 24.6 ounces (678 to 697 grams) from 1955 to 1962, the length of carcasses increased from 36.9 to 37.8 inches (93.8 to 95.9 centimetres), and the backfat thickness decreased from 1.28 to 1.10 inches (3.26 to 2.80 centimetres). The modern broiler chick is an example of the success obtained by crosses between breeds that have been specialized for different lines of production without close inbreeding.

Limitation to selection

When there is a considerable amount of hereditary variation, it is possible to change a breeding population considerably in about five to 10 generations of intense selection. Sooner or later, the response to selection decreases, and ultimately a selection limit is reached. This may be due to an exhaustion of the genetic variation or, more likely, to a disturbed gene balance, especially if the selection is concentrated on one, or a very few, traits only, without considering the fertility and viability of the animals. The remedy in such a situation is either a cautious introduction of new genes from another population or deliberate cross-breeding to increase the genetic variation. In either case, selection in the gene-enriched population can start again. Another possibility is to relax the selection for a number of generations, giving the population time to recover, but this is a rather time-consuming process. (I.J.)

Livestock research and development. In the developed countries the trend toward feeding ruminant livestock with feed grains and forages grown on arable land is causing concern as an inefficient utilization of land in the face of the increasing need for grain as human food. More research effort, therefore, is likely to be directed toward increasing the yield of animal products from rangeland and permanent pasture. Large-scale nitrogen fertilization of pastures and other forages will improve their yield and nutritive value, and the practice will become justified as new forages are developed and the cost of nitrogen is lowered; in fact, world fertilizer consumption increased by some 75 percent in the 1970s. An alternative would consist of adding urea and ammonia salts to carbohydrates and forages as the sole source of nitrogen. That would increase the bacterial synthesis of protein in the rumen of cattle and sheep. Much urea is now fed to livestock; if its price trends low enough, it could be fed with forage whose protein content would thus be of less importance than its digestibility. Another possibility lies in the chemical treatment of high-lignin forages to increase digestibility. These techniques offer some hope for reduced consumption of grain by ruminants.

Large-scale fertilization of pastures

Dairy cows. The full potential of the dairy cow has not yet been realized. Worldwide, milk production per cow increased by 3 or 4 percent per year for more than 10 years until the early 1970s—largely from genetic improvements, better nutrition, and improved management—but has increased at a slower rate since. The effects of physiological and genetic limitations will probably not be felt until the year 2000. Meanwhile, breeding research and adoption of controlled environment housing will continue to increase milk production.

Beef cattle and swine. Productivity of beef cattle and swine is related to reproductive capacity, rate of growth, and the lean-meat percentage of the carcass. Of these factors, low reproduction rate is the most important. The percentage of mated beef cows who raise calves to maturity is 75 to 80 percent in the United States; 90 percent for The Netherlands and Denmark; but only 30 to 50 percent in developing countries. In the future, selection programs designed to increase the percentage of multiple births in beef cattle and swine will help to increase production efficiency. Reduction in losses both before and after birth can be realized by additional research. Progress in breeding and selection of swine for lean-meat carcasses has been achieved, a process that should be done also with beef cattle, because market surveys have shown preference for lean Holstein beef rather than the usual English breeds. The preference of beef feeders for crossbreeds is already well-established; heterosis (hybrid vigour) often results in a 5 to 20 percent increase over the best parental breed. Research on climate control at high temperature should result in improved production efficiency.

Efficiency in poultry production

Poultry. The efficiency of poultry production has undergone marked improvement. The potential limit for the laying hen is one two-ounce (56-gram) egg from each 2½-pound (1.1-kilogram) hen daily for one year, based on three pounds (1.4 kilograms) of feed per dozen eggs. The average hen, however, lays only about 230 eggs per year, weighs four pounds (1.8 kilograms), and requires four pounds of feed to produce a dozen eggs. One point of potential gain lies in causing the hen to lay eggs without shells, although public acceptance might be limited. The ultimate for feed conversion in the broiler is one pound (0.45 kilograms) of meat for every pound of feed utilized in a five-week growing period. The present conversion ratio is 2.2 pounds (one kilogram) feed for every pound of meat, utilized over a seven- to nine-week growing period. Increased productivity in poultry of the future will result from new genetic combinations and further improvements in environmental and management factors. One promising area for future research may be in exploration of performance as related to environmental conditions.

The future for livestock. The biological limits in livestock productivity have not been achieved in any class of farm animal. Fundamental research in biochemistry, endocrinology, environmental physiology, genetics, parasitology, microbiology, and disease has provided knowledge that will make progress possible. For example, swine have reached market weights of over 200 pounds (90 kilograms) in 120 days with little more than two pounds of feed per pound of gain. Some hogs have yielded loin eyes of nine square inches (58 square centimetres) and 50 percent of the carcass weight in ham and loin. Sows have farrowed more than 30 pigs in one litter. Knowledge of the endocrine system and its hormone production now makes it possible to synchronize estrus (heat), cause superovulation, and ultimately to control body composition. Fibre digestion efficiency in ruminants depends on the type of bacteria and protozoa in the rumen. By controlling these microbes, the quality of protein synthesized can be improved. There is reason to hope that in the fairly near future hogs will be marketed 100 days from birth and beef animals within six months; that each hen will produce two eggs a day; and that the beef cow will produce two calves per year. (R.E.S.)

COOPERATIVE AND GOVERNMENTAL PROMOTION OF ANIMAL BREEDING

In most countries in which animal breeding has reached a fairly high level, associations were formed long ago (beginning in Great Britain in the late 18th century) with the aim of promoting the development of existing breeds. At first the primary objectives of breeders' associations were to publish herdbooks and to lay down rules for registration of "purebred" animals, to arrange shows and fairs, and to work for the dissemination of the breeds at home and abroad. Later, these associations started performance testing or worked in close cooperation with other organizations developed for this purpose. In European cattle breeding there is a trend toward concentration into one organization of the various activities, such as performance and progeny testing, artificial insemination, and other services, among which herdbook registration may be only a minor detail. A similar trend has appeared also in the breeding of sheep, pigs, and fur animals. With regard to the breeding of animals for sport or as pets, traditional breeders' associations are, and probably will continue to be, of great importance.

Breeders' associations

Scientific research has been the foundation, and education the impelling force, in the accelerated development of animal breeding in the 20th century. Specialized research institutes and experimental stations, with cooperating agricultural colleges and extension divisions, have arisen in almost every country in which animal production is an important enterprise. Most information based on laboratory research and breeding experiments has arisen in the United States and Great Britain, but Denmark seems to have been the first country to organize an agricultural extension service that reached all levels of farming and animal husbandry.

Organizations have been established in several countries to promote animal science and production; for example, in the United States, the American Society of Animal Science and the American Dairy Science Association; in Great Britain, the British Society of Animal Production; and in Germany, Deutsche Gesellschaft für Züchtungskunde. In 1966 these national organizations formed the World Association for Animal Production, the major objective of which is to arrange periodic world conferences for the exchange and dissemination of knowledge in the field of animal science. (I.J.)

Animal nutrition

Animals in general require the same nutrients as humans. (This section is concerned primarily with feed for livestock and poultry.) Some feeds, such as pasture grasses, hay and silage crops, and certain cereal grains, are grown specifically for animals. Other feeds, such as sugar-beet pulp, brewers' grains, and pineapple bran, are by-products remaining after a food crop has been processed for human use. Surplus food crops, such as wheat, other cereals, fruits, vegetables, and roots, may also be fed to animals. In this way such surpluses are converted into meat, milk, and eggs for the human diet.

History does not record when dried roughage or other stored feeds were first given to animals. Most early records refer to nomadic people who, with their herds and flocks, followed the natural feed supplies. When animals were domesticated and used for work in crop production, some of the residues were doubtless fed to them.

Preservation of green forages such as beet leaves and corn (maize) plants by packing them in pits in the earth has long been practiced in northern Europe. The idea of making silage as a means of preserving and utilizing more of the corn plant was gradually developed in Europe and brought from France to the United States in the 1870s. When the mature, dried corn plant was fed to cattle in the winter much of the coarse stem was wasted, but when it was chopped and ensiled (made into silage) everything was eaten. The first effort to evaluate feeds for animals on a comparative basis was apparently made by Albrecht Thaer (1752–1828), in Germany, who developed "hay values" as measures of nutritive value of feeds. Tables of the value of feeds and of the requirements of animals in Germany followed and were later used in other countries. Present-day knowledge represents an expansion and improvement of these early efforts.

BASIC NUTRIENTS AND ADDITIVES

The basic nutrients that animals require for growth, reproduction, and good health include carbohydrates, protein, fat, minerals, and vitamins. The energy needed for growth and activity is derived primarily from carbohydrates and fats. Protein may also supply energy, however, if other sources are not adequate or if it is supplied in great excess above the needs of the body.

Animals need a source of energy to sustain life processes within the body and for muscular activity. When the energy intake of an animal exceeds its requirements, the surplus is stored as body fat, which can be utilized later if the energy consumed in food is not adequate.

Proteins. All animals require a small amount of protein for the daily repair of muscles, internal organs, and other body tissues. For immature animals, protein is also needed for growth of the muscles and other parts of the body. Since milk, eggs, and wool contain much protein, additional amounts are needed in the food of animals producing these.

Proteins are composed of more than 20 different amino acids, which are liberated during digestion. Animals with simple stomachs, including humans, monkeys, swine, poultry, rabbits, and mink, require correct amounts of the following 10 essential amino acids daily: arginine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, trytophan, and valine. In addition to these, poultry need glycine and glutamic acid for growth. High-quality protein supplied by eggs, milk, fish meal, meat by-products, and soybean meal contains correct amounts of the essential amino acids. Poor-quality protein, such as that in corn grain (maize), contains too little of one or more essential amino acids. Feeds having poor-quality proteins are made useful by blending them with other feeds that supply the lacking amino acids.

Role of ruminants

The quality of the protein in the food is of little importance to ruminants, including cattle, sheep, goats, and the other animals that have four stomachs, because the bacteria that aid in the digestion of food in the rumen (first stomach) use simple nitrogen compounds to build proteins in their cells. Further on in the digestive tract the animals digest the bacteria. By this indirect means, ruminants produce high-quality protein from a food that might originally have contained poor protein, or from urea (a nitrogen compound). Very young ruminants, such as calves, lambs, and kids, however, need good-quality protein until the rumen develops sufficiently for this bacterial process to become established.

Carbohydrates and fats. Most animals get energy from carbohydrates and fats, which are oxidized in the body. These yield heat, which maintains body temperature, furnishes energy for growth and muscle activity, and sustains vital functions. Animals need much more energy (and more total feed) for growth, fattening, work, or milk production than simply for maintenance.

The less complex carbohydrates (sugars and starches) are readily digested by all animals. The complex carbohydrates (cellulose, hemicelluloses) that make up the fibrous stems of plants are broken down by bacterial action in the rumen of cattle and sheep or in the cecum of rabbits and horses. Such complex carbohydrates cannot be digested by men, or, to any appreciable extent, by dogs, cats, birds, or laboratory animals. Thus, ruminants and some herbivorous animals obtain much more of the energy-giving nutrients from the carbohydrates of plants than the simple-stomached carnivores and omnivores, for which fibrous materials have little or no energy value.

Fat in feeds has a high nutritive value because it is highly digestible, and because it supplies about two and one-quarter times as much energy as starch or sugar. While fat has high nutritive value, it can be replaced by an equivalent amount of digestible carbohydrates in the food, except for small amounts of essential fatty acids. Very small amounts of the unsaturated fatty acid linoleic, contained in some fats, are necessary for growth and health. Usual animal feeds, however, supply ample amounts of this acid unless it has been removed by processing.

Minerals. Minerals essential for animal life include common salt (sodium chloride), calcium, phosphorus, sulfur, potassium, magnesium, manganese, iron, copper, cobalt, iodine, zinc, molybdenum, and selenium. The last six of these are poisonous to animals if excessive amounts are eaten.

All farm animals generally need more common salt than is contained in their feeds, and they are supplied with it regularly. Of the other essential minerals, phosphorus and calcium are most apt to be lacking, because they are heavily drawn upon to produce bones, milk, and egg shells. Good phosphorus supplements are bone meal, dicalcium phosphate, and defluorinated phosphates. Egg shells are nearly pure calcium carbonate. Calcium may readily be supplied by ground limestone, ground shells, or marl that is high in calcium.

Losses caused by goitre

Small amounts of iodine are needed by animals for the formation of thyroxine, a compound containing iodine, secreted by the thyroid gland. A serious deficiency of iodine may cause goitre, a disease in which the thyroid gland enlarges greatly. In certain regions goitre has caused heavy losses of newborn pigs, lambs, kids, calves, and foals. Goitre can be prevented by supplying small amounts of iodized salt to the mother before the young are born.

In some areas, soil and forage are deficient in copper and cobalt, which are needed along with iron for the formation of hemoglobin, the oxygen-carrying pigment of the red blood cells. In these areas, farm animals may suffer from anemia unless the deficiency is corrected by means of a suitable mineral supplement.

Iron, used in hemoglobin formation, is amply supplied in most animal feeds, except milk. The only practical problem with iron deficiency occurs in young suckling pigs before they start to consume other feeds in addition to milk.

Though manganese is essential for animals, the usual rations for all farm animals, except poultry, supply sufficient quantities. A lack of manganese may cause the nutritional disease of chicks and young turkeys called slipped tendon (perosis) and also may cause failure of eggs to hatch. Normal rations for swine are often deficient in zinc, especially in the presence of excess calcium. Adding 100 parts per million of zinc carbonate cures zinc-deficiency symptoms, which include retarded growth rate and severe scaliness and cracking of the skin (parakeratosis). A trace of selenium is necessary for normal health of animals; excessive amounts found in forages in some regions poison animals and may cause death. To furnish both calcium and phosphorus, livestock may be allowed free access to such a mixture as 60 percent dicalcium phosphate and 40 percent common salt. Trace mineralized salt is used when copper or cobalt may be deficient. Animals are given access to common salt separately, so they will not be forced to eat more of the other minerals than they require to get the amount of salt they need.

Vitamins. Known vitamins include A, C, D, E, and K, and the B group: thiamine, riboflavin, niacin, pantothenic acid, choline, biotin, folic acid, and vitamins B_6 and B_{12}.

Vitamin A, the one most apt to be lacking in livestock feeds, is required for growth, reproduction, milk production, and the maintenance of normal resistance to respiratory infections. All green-growing crops are rich in carotene, which animals can convert into vitamin A. Vitamin A supplement is added to animal rations to ensure a supply when they are not on good pasture.

Vitamin D in hay

Vitamin D enables animals to use calcium and phosphorus; a deficiency causes rickets in young growing animals. The ultraviolet rays of sunlight produce vitamin D from the provitamin in the skin. Field curing of hay develops vitamin D through the action of the sunlight on ergosterol in the hay crops. Certain fish oils are very rich in vitamin D. Livestock that are outdoors in the sunlight much of the time have a plentiful supply of vitamin D. Under winter conditions in cold regions, cattle, sheep, and horses ordinarily get ample amounts from the hay they are fed; laboratory animals may be deficient unless a supplement is added.

The vitamin B group is not important in the feeding of cattle, sheep, and other ruminants because the bacteria in the rumen synthesize these vitamins. Very young calves, poultry, swine, and other simple-stomached animals require the B vitamins in their diets. Of these, riboflavin, niacin, pantothenic acid, and vitamin B_{12} are most likely to be deficient in ordinary feeds; special supplements are needed by pigs, poultry, and laboratory animals. Choline may also be deficient in poultry feeds.

Vitamin E is necessary for normal hatching of eggs. It plays a role along with selenium in preventing muscle stiffness and paralysis (dystrophy) in lambs, calves, and chicks under certain conditions. Vitamin C, which prevents scurvy in humans and guinea pigs, can be synthesized in the bodies of other animals and need not be supplied in their food. Vitamin K is synthesized by bacteria in the intestinal tract; a dietary supply is usually not important.

Antibiotics and other growth stimulants. Since about 1950, antibiotics have been added to diets of growing animals because they increase the rate of growth and decrease death loss. Those most commonly used are chlortetracycline, oxytetracycline, bacitracin, and penicillin. An antibiotic helps to overcome the growth-depressing effects of an inadequate diet or of imperfect management practices, but its effectiveness differs among animal species.

Diethylstilbestrol, testosterone, progesterone, estradiol, and dienestrol in various combinations are administered orally or subcutaneously to stimulate growth of cattle and sheep for meat production.

COMPOSITION AND VALUATION OF FEEDS

The usual chemical analyses of feeds provide information on the amount of dry matter, protein, fat, fibre, and ash contained in the feed. Energy value, mineral elements, and vitamins are also determined; these values are included in complete tables of feed composition.

Animal protein requirements

Determination. Digestion and balance experiments measure the degree to which the various components of a feed are absorbed and retained by the animal body. Protein requirements are expressed as the amounts of digestible protein needed for growth or other body functions. The amounts of energy needed are measured as digestible energy (DE), metabolizable energy (ME), net energy (NE), or total digestible nutrients (TDN). The total gross energy value of a feed is the amount of heat liberated when it is burned in a bomb calorimeter. Only a part of this energy is available to the animal because some passes through the body without being digested. Furthermore, some of the DE is not utilized but is excreted in the urine as urea. There is a further energy loss from the ME as heat of fermentation and as gas produced by bacteria in the digestive tract. This loss is appreciably greater in ruminants than in pigs, chickens, or other nonruminants. The work of eating, digesting, and metabolizing food may also be subtracted from the food energy, resulting in the NE, or useful energy value of a food. The TDN value of a feed represents the sum of the digestible protein, digestible fat × 2.25, digestible nitrogen-free extract, and digestible fibre. This measure is less useful than NE because it considers neither the fermentation and heat losses during digestion and metabolism nor the fact that energy is utilized more efficiently for maintenance or milk production than for growth and fattening.

Optimization of nutrient-cost ratio. Feed costs vary widely from season to season; it is often possible for producers to realize substantial savings by wise selection of the feed ingredients used to formulate complete rations. It is much easier for large commercial feed companies with operations in different regions to take advantage of low costs than it is for individual relatively small-scale livestock producers.

Least-cost programming of feed mixtures makes it possible to use electronic computers for selecting the correct amounts of the lowest-cost feed ingredients that will fully satisfy the nutrient requirements of a specific type of animal at a particular stage of development. The system has successfully formulated highly palatable rations that yield maximum production at lowest cost. As in other areas of agriculture, the large-scale producer, with the ability to purchase and operate advanced systems, has an economic advantage.

BASIC TYPES OF FEEDS

Animal feeds are classified as follows: (1) concentrates, high in energy value, including (*a*) cereal grains and their by-products (barley, corn [maize], oats, rye, wheat), (*b*) high-protein oil meals or cakes (soybean, cottonseed, peanut [groundnut]), (*c*) by-products from processing of sugar beets, sugarcane, and (*d*) animal and fish by-products; (2) roughages, including (*a*) pasture grasses, (*b*) hays, (*c*) silage, (*d*) root crops, and (*e*) straw, stover (stalks). The composition of a few of the most commonly used feeds is shown in Tables 1 and 2.

Concentrate foods. *Cereal grains and their by-products.* In the agricultural practices of North America and northern Europe, barley, corn (maize), oats, rye, and sorghums are grown almost entirely as animal feed, although small quantities are processed for human consumption as well. These grains are fed, whole or ground, either singly or mixed with high-protein oil meals or other by-products, minerals, and vitamins, to form a complete feed for pigs and poultry or an adequate dietary supplement for ruminants and horses.

The production of grains is seasonal because of temperature or moisture conditions or a combination of both. It is necessary to produce a full year's supply during the limited growing season. The grain is dried to prevent sprouting or molding; the grain is then stored in containers or buildings where insects and rodents cannot destroy it. It is generally desirable to store more than a year's supply of the grains to be used as feed, because crop failures sometimes occur.

Grainfeed storage

Table 1: Concentrates Commonly Used as Animal Feeds

type of concentrate	composition (%)					
	dry matter (total)	protein	fat	fibre	ash	NFE*
Barley	88.6	12.0	2.1	7.4	3.3	63.8
Beans (*Phaseolus* species)	89.0	23.0	1.2	4.1	3.9	56.8
Bone meal, steamed	95.0	12.1	3.2	2.0	71.8	5.9
Brewers' grains	92.0	25.9	6.2	15.0	3.6	41.3
Brewers' yeast	93.0	44.6	1.1	3.0	6.4	37.9
Citrus pulp	90.0	6.6	4.6	13.0	6.0	59.8
Coconut meal	93.0	20.4	6.6	12.0	6.9	47.1
Corn	86.5	9.1	4.0	1.8	1.2	70.4
Corn distillers' grains	91.6	29.1	8.9	11.5	3.1	39.0
Corn ears (corn and cob meal)	87.0	8.1	3.2	8.0	1.6	66.1
Corn gluten feed	90.0	25.3	2.4	8.0	6.3	48.0
Cottonseed, whole	92.7	23.1	22.9	16.9	3.5	26.3
Cottonseed meal	94.0	41.0	4.3	12.0	6.2	30.5
Fish meal	92.0	63.2	4.4	—	21.7	27.0
Hominy feed	90.6	10.7	6.5	5.0	2.5	65.9
Linseed meal	91.0	35.3	5.2	9.0	5.6	35.9
Meat and bone meal	94.0	50.6	9.5	—	29.1	48.0
Milk, dried skimmed	94.0	33.5	0.9	—	7.6	26.3
Molasses, cane	75.0	3.2	0.1	—	8.1	63.6
Oats	91.0	12.1	4.8	12.0	3.2	58.9
Palm-kernel meal	91.4	19.2	6.7	11.9	3.9	49.7
Peanut meal	92.0	47.4	1.2	13.0	4.5	25.9
Pineapple bran or pulp	88.6	4.2	1.6	18.4	2.6	61.8
Poultry waste	93.1	52.1	13.6	—	8.5	18.9
Rapeseed meal	93.6	37.1	6.3	13.7	6.0	30.5
Rice, grain	89.0	7.3	1.9	9.0	4.5	66.3
Rice bran	91.0	13.5	15.1	11.0	10.9	40.5
Rye	89.0	11.9	1.6	2.0	1.7	71.8
Safflower meal	91.0	19.7	6.0	31.0	3.7	30.6
Sesameseed meal	93.0	47.9	5.1	5.0	9.3	25.7
Sorghum	88.0	10.3	2.6	2.1	1.9	71.1
Soybean hulls	91.3	12.5	2.1	35.5	4.6	36.6
Soybean meal	89.8	50.9	0.8	2.8	5.6	29.7
Soybean seed	90.0	37.9	18.0	5.0	4.6	24.5
Sugar-beet pulp	91.0	9.1	0.6	19.0	3.6	58.7
Wheat	89.0	12.7	1.7	3.0	1.6	70.0
Wheat bran	89.0	16.0	4.1	10.0	6.1	52.8
Wheat-germ meal	90.0	26.2	10.9	3.0	4.3	45.6
Wheat middlings	89.0	18.0	3.6	2.0	2.5	62.9
Whey, dried	94.0	13.8	0.8	—	9.7	69.7

*NFE = nitrogen-free extract.

Source: Data selected from U.S. National Academy of Sciences, National Research Council, *United States–Canadian Tables of Feed Composition*, 1969.

High-protein meals. Vegetable seeds produced primarily as a source of oil for human food and industrial uses include soybeans, peanuts (groundnuts), flaxseed, cottonseed, coconuts, oil palm, and sunflower seeds. After these seeds are processed to remove the oil, the residues, which may contain from 5 percent to less than 1 percent of fat and 20 to 50 percent of protein, are marketed as animal feeds. Cottonseed and peanuts have woody hulls or shells, which are generally removed before processing—if the hulls or shells are left intact, the resulting by-product is higher in fibre and appreciably lower in protein and energy value.

These feeds are used as supplements to roughages or cereal grains and other low-protein feeds to furnish the protein needed for efficient growth or production. The supplement chosen for a particular ration depends largely upon the cost and availability of supply.

By-products of sugar beets, sugarcane. From the sugar-beet industry come beet tops, which are used on the farm either fresh or ensiled, and dried beet pulp and beet molasses, which are produced in the sugar factory. Cane molasses is a residue from cane-sugar manufacture. These are all palatable, high-quality sources of carbohydrates. Sugarcane bagasse (stalk residue) is fibrous, hard to digest, and of very low feed value. In some European countries fodder beets and some other roots are grown as animal feed. Citrus molasses and dried citrus pulp, which are generally available at low cost as by-products of the citrus-juice industry, are often used as high-quality feeds for cattle and sheep.

Other by-product feeds. By-products or residues from commercial processing of the cereal grains to produce human food supply large quantities of animal feeds. The largest group of these comprises feeds derived from the milling of wheat in the production of flour including wheat bran, wheat middlings, wheat-germ meal, and wheat-mill feed. In some areas bakery wastes, such as stale and left-over bread, rolls, and various pastry products, are ground and used as filler or feed for pets and farm animals. Rice bran and rice hulls are obtained in similar fashion from the mills that polish rice for human food. Corn gluten feed, corn gluten meal, and hominy feed are produced as by-products from the manufacture of starch for industrial and food uses.

Brewers' grains, corn-distillers' grains, and brewers' yeast are useful animal feeds and are collected from the dried residues of the fermentation industries that produce beer and distilled spirits. Waste products from pineapple-canning plants include pineapple bran or pulp and the ensiled leaves from the plant. By-products from the abattoirs and meat-packing plants that process animals into meat include such feeds as meat and bone meal, tankage, meat scraps, blood meal, poultry waste, and feather meal. Dried skim milk, dried whey, and dried buttermilk are feed by-products from the dairy industry. Various types and qualities of fish meals are produced by the fish-processing plants. These feeds contain 50 percent or more of high-quality protein and the mineral elements calcium and phosphorus. Steamed bone meal is particularly high in these important minerals.

Bone and fish meals

Roughages. *Pasture.* Pasture grasses and legumes, both native and cultivated, are the most important single source of feed for cattle, horses, sheep, and goats. During the growing season they furnish most of the feed for these animals at a cost lower than for feeds that need to be harvested, processed, and transported. Hundreds of different grasses, legumes, bushes, and trees are acceptable as feeds for grazing animals. The nutritive value of the cultivated varieties has been studied, but information is incomplete for many of those that occur naturally. The composition of some of the forage crops most widely used is shown in Table 2.

Hay. Hay is produced by drying grasses or legumes when they approach the stage of maximum plant growth and before the seed develops. This stage has been shown to give maximum yields of digestible protein and carbohydrates per unit of land area. The moisture content must be reduced to 22 percent or less to prevent molding, heating, and spoilage during storage. Legume hays, such as alfalfa and clovers, are high in protein, while the grasses (such as timothy and Sudan grass) are lower in protein but vary considerably depending upon the stage of maturity and level of nitrogen fertilization applied to the crop. Stored hay is fed to animals when sufficient fresh pasture grass is not available.

Legume and grass hays

Silage. Silage is made by packing immature plants in a storage container to exclude the air and allow fermentation to develop acetic and other acids, which preserve the moist feed. Storage may be in upright tower silos or in trenches in the ground. Best quality silage results when the forage is ensiled with a moisture content of 50 to 65 percent. Lower moisture levels can cause difficulty in obtaining sufficient packing to exclude air and may result in molding or other spoilage. Too high a moisture content causes nutrient losses by seepage and results in the production of excessively acid, unpalatable silage. Ensiled forage can be stored for a longer period of time with lower loss of nutrients than dry hay. The nutritive value of silage depends upon the type of forage ensiled and how successfully it has been cured. Corn (maize), sorghums, and grasses, and sometimes leguminous forages, are used in making silage.

Root crops. These are used less extensively as animal feed than was true in the past, for economic reasons. Mangels, rutabagas, cassava, and sometimes surplus potatoes are used as feed. As shown in Table 2, they are lower in dry-matter content than are most of the other feeds listed. They are relatively low in protein also and provide mostly energy.

Table 2: Roughages Commonly Used as Animal Feeds

type of roughage	composition (%)					
	dry matter (total)	protein	fat	fibre	ash	NFE*
Dry roughages						
Alfalfa hay	90.0	16.6	2.0	26.8	8.5	36.1
Berseem	90.6	13.4	3.1	24.4	11.0	38.7
Bromegrass hay	92.8	5.4	2.8	31.7	7.6	45.3
Clover hay, crimson (*Trifolium* species)	87.4	14.8	2.0	28.1	8.3	34.2
Corn cobs	90.4	2.5	0.5	32.4	1.5	65.2
Corn fodder	87.2	5.1	1.0	32.4	6.2	42.5
Cottonseed hulls	90.3	3.9	1.4	42.9	2.5	39.6
Lespedeza hay	93.0	14.6	3.7	28.6	5.5	40.6
Oat hay	88.2	8.1	2.7	27.3	6.6	43.5
Oat straw	90.1	4.0	1.9	36.9	7.4	39.9
Peanut hay, without nuts	90.6	10.0	3.2	23.6	9.6	44.2
Rice hulls	92.0	3.0	0.8	40.7	19.1	28.4
Rice straw	92.5	3.9	1.4	33.5	14.5	39.2
Sorghum, without heads	85.1	4.5	1.8	27.7	8.2	42.9
Soybean hay	89.2	14.5	2.7	28.6	7.2	36.2
Timothy hay	87.7	7.6	2.3	29.1	5.4	43.3
Wheat straw	92.6	3.9	1.5	37.0	8.3	41.9
Green roughages						
Alfalfa	21.1	4.3	0.5	5.5	2.0	8.8
Barley pasture	20.0	5.2	0.8	3.7	3.3	7.0
Beet tops	20.7	2.6	0.4	2.8	7.5	7.4
Bermuda grass	36.7	4.2	0.8	9.5	3.8	26.0
Berseem	18.8	2.7	0.7	4.4	2.8	8.2
Bluegrass	30.5	5.3	1.1	7.6	2.2	14.3
Bromegrass	32.5	7.2	1.4	7.3	3.4	13.2
Clover, ladino	18.7	4.7	0.9	2.7	2.1	8.3
Fescue	23.9	2.8	0.8	7.1	2.0	11.2
Guinea grass	26.8	1.4	0.4	11.5	3.0	10.5
Lespedeza	31.1	5.8	0.8	8.5	3.3	12.7
Mangels, beets	10.6	1.4	0.1	0.9	1.1	7.1
Molasses grass	30.9	1.3	0.9	16.0	2.8	9.9
Napier grass	22.0	1.1	0.3	9.0	2.6	11.2
Orchard grass	23.8	4.4	1.1	5.6	2.7	10.0
Para grass	27.8	1.8	0.4	10.0	2.9	12.7
Rape	16.3	2.9	0.6	2.6	2.2	8.0
Rye grass	24.3	4.0	1.0	5.3	3.8	10.2
Sudan grass	17.6	3.0	0.7	5.4	1.6	6.9
Sugarcane	23.2	1.0	0.8	6.8	1.2	13.4
Wheatgrass (*Agropyron* species)	30.8	7.3	1.1	6.8	3.3	12.3
Silages						
Alfalfa, wilted	36.2	6.4	1.2	10.9	3.1	14.6
Clover	39.9	7.9	1.3	12.7	4.0	14.0
Corn	25.6	2.2	0.8	6.6	1.5	14.5
Pineapple tops	21.3	1.6	0.6	4.8	1.5	12.8
Sorghum	28.9	2.3	0.8	7.8	2.2	15.8
Soybean	28.0	4.1	0.9	8.7	2.8	11.5
Timothy	40.8	4.4	1.3	14.2	2.9	17.0
Root crops						
Beets	13.0	1.6	0.1	0.9	1.5	8.9
Carrots (*Daucus* species)	11.9	1.2	0.2	1.1	1.2	8.2
Cassava	32.6	1.1	0.3	1.4	1.0	28.8
Potatoes	24.6	2.2	0.1	0.5	0.9	20.9
Turnip (*Brassica rapa*)	9.3	1.3	0.2	1.1	0.9	5.8

*NFE = nitrogen-free extract.

Source: Data selected from U.S. National Academy of Sciences, National Research Council, *United States–Canadian Tables of Feed Composition*, 1969.

Straw and hulls. Quantities of straws remaining after wheat, oats, barley, and rice crops have been harvested are used as feed for cattle and other ruminants. The straws are low in protein and very high in fibre; digestibility is low. Straw is useful in maintaining mature animals during periods of shortage of other feeds, but it is too low in quality to be satisfactory for long periods without being supplemented with other feeds to supply the protein, digestible energy, and minerals needed for growth and production. Treatment of straw with alkali markedly increases the digestibility of the cellulose, augmenting its value as a source of energy for animals.

Corncobs, corn stalks, cottonseed hulls, and rice hulls can also be used as sources of fibre in ruminant rations. Rice hulls are lower in value, while the others are similar to straw.

(J.K.L.)

CROP-FARMING

Cereal farming

Cereals, or grains, are members of the grass family cultivated primarily for their starchy seeds (technically, dry fruits), which are used for human food, livestock feed, and as a source of industrial starch. Wheat, rice, corn (maize), rye, oats, barley, sorghum, and some of the millets are common cereals.

The cultivation of cereals varies widely in different countries and depends partly upon the degree of economic development. The condition and purity of the seed is receiving increasing attention. Other factors include the nature of the soil, the amount of rainfall, and the techniques applied to promote growth. In illustrating production problems, this section will use wheat as the example. For information on the cultivation of other cereal crops such as rice, see articles on the individual crops in the *Micropædia.* For information on the food value and processing of cereals, see the article FOOD PROCESSING.

CULTIVATION OF WHEAT

Wheat can be cultivated over a wide range of soils and can be successfully grown over large portions of the world, ranging in altitude from sea level to over 10,000 feet. Annual rainfall of 10 inches (254 millimetres) is generally considered the minimum, and the soil should be sufficiently fertile. (Barley can be grown in soil less fertile than that required for wheat.) Soil benefits from a good humus content (partially decayed organic matter), and chemical fertilizers are also helpful.

Importance of seed purity

Purity of the seed is important. The seed wheat (or other cereal seeds) must be true to its particular variety and as free as possible from foreign seeds. Seeds are frequently cleaned to avoid contamination by other seed crops. Modern cleaning methods employ such devices as oscillating sieves or revolving cylinders. Seed obtained with a combine harvester is often unsuited for use as seed wheat without preliminary treatment. Spring and winter varieties exist for both wheat and barley. Winter varieties generally produce better crops. Winter wheat should form a good root system, and the plant should begin to form new shoots before the cold weather sets in; winter wheat is likely to have more tillers than spring wheat.

The rate of sowing varies from 20 pounds per acre (22.5 kilograms per hectare) upward. Depth of sowing, usually one to three inches (2.5 to 7.5 centimetres), can be less in certain areas.

Breeding. Wheat and other cereals are self-fertilized. The pollen carried by the stamen of a given flower impregnates the pistil (stigma and ovary) of the same flower, enabling the variety to breed true. Wheat flowers are grouped in spikelets, each bearing from two to nine flowers, or florets. To produce new varieties by cross-fertilization, the cereal breeder artificially transfers the stamen from one variety to the flower of another before self-fertilization takes place. The production of a sufficient supply of the new type of seeds for sowing is time-consuming and expensive, but it allows new varieties to be evolved, retaining the desirable characteristics from each parent. For example, especially in the United Kingdom and Australia, varieties of the wheat that yield well often produce flour of poor baking quality; proper selection of parent plants permits new varieties to be produced that yield well and still possess good baking qualities.

Other reasons for developing new varieties include resistance to rust (fungus; see below) and other diseases, resistance to drought, and development of stronger and shorter straw to make harvesting easier.

Seedbed preparation. Various types of plowing machinery and other implements are employed to render the soil more suitable for seed wheat planting. The equipment used depends upon such factors as the climate, the nature of the ground, and the rainfall. Tillage is the process of preparing soil for cultivation purposes. The practices used and the implements employed vary considerably. Serious soil erosion may require special procedures to maintain clods and plant residues in the soil.

Tillage

In North America it is normal practice to grow wheat on the same ground for as long as sufficiently clean crops are produced, but eventually the ground must rest fallow for a year. The moisture of the land at the time of sowing is an important factor. The ancient procedure of growing legumes occasionally to improve the soil is still common in Europe, though less so in North America. Fertilization of the ground is useful to increase the crop yield, but it does not generally increase the protein content of the crop. In the large collective state farms of the Soviet Union, huge harrows set with spikes or teeth are employed, as well as the disk cultivating plow set with disks that break up the soil; the scarifier, a machine that pulverizes the soil, is popular in Australia.

PLANT PROTECTION

Winter crops are frequently disturbed by frost, and the ground must then be rolled in the spring to consolidate the soil around the roots. If soil has become crusted by heavy rains followed by surface drying, the crop is usually harrowed in the spring to aerate the soil and kill young weeds. Although all of the required mineral nutrients may be added to the soil at the time of sowing, sometimes only part of the nitrogenous fertilizers is added at that time, and the remainder is applied to the growing crop in the form of a top dressing. In the cultivation of spring wheat all of the fertilizer is usually added before or at sowing time, but sometimes a small portion is reserved for later.

Weeds. Weeds present difficulties, as they compete with cereal crops for water, light, and mineral nutrients. The infestation of annual seeds planted in a field may cause many weeds in that field for successive years. Charlock or wild mustard, wild oats, crouch grass, and other common weeds are disseminated by wind, water, and birds.

Insects. In addition to weeds, wheat and other cereals are seriously affected by insects.

Grasshoppers and locusts

Grasshoppers and locusts cause immense damage. Spraying from airplanes with chemicals such as gamma BHC, Dieldrin, chlordane, or Toxaphene is effective; on small farms grasshopper control is often accomplished by weed killers such as MCPA (2-methyl-4-chlorophenoxyacetic acid) and 2,4-D (2,4-dichlorophenoxyacetic acid).

The eggs of click beetles are laid in the soil, and the larvae, called wireworms, live underground for some years, feeding on the roots and stems of the young plants (particularly slow-growing plants). To combat such damage, chemical seed dressing is used together with nitrogenous fertilizers. Other measures use such chemicals as gamma BHC (Lindane) or Dieldrin.

Aphids attack many plants, and the wheat aphid, or greenbug, causes damage throughout the world. Preventive action includes preparing a good seedbed, sufficient fertilization, and early sowing.

The wheat stem sawfly (*Cephus cinctus*) is found in many parts of the world. Infested wheat shows fallen straw filled with a fine sawdust material harbouring brown-headed larvae that pass the winter in the base of the wheat straw; the wasplike adult insect emerges around June. The females thrust their eggs into the upper plant tissues, and the larvae feed within the stem toward the base until the stem collapses. Varieties of Manitoba wheat such as Rescue and Chinook are reasonably resistant to the pest, and thorough plowing in of the infested stubble is generally effective. Certain crops, such as brome grass, attract this pest and may be grown on the borders of wheat crops to distract the pests away from the wheat. The hessian fly (*Mayetiola,* or *Phytophaga, destructor*), resembling the mosquito, attacks the stems of wheat, barley, and rye. Late wheat usually escapes damage from this pest.

Many wheats in central Europe and the Middle East have shown evidence of attacks from the wheat bug (*Weizenwanze,* or *blé punaisé*). The two main varieties are the *Aelia* and the *Eurygaster.* The eggs are laid in the spring, and the new generation appears in the summer. When the wheat is harvested, the bugs leave the stubble field and migrate to nearby foliage for the winter. To thrive and multiply, wheat bugs require sun, warmth, and absence of pronounced dampness.

Gluten damage

The wheat bugs puncture the grain and introduce by means of their saliva an enzyme that profoundly modifies the nature of the gluten. The puncture mark can be seen on the grain, usually surrounded by a yellow patch, and sometimes the grain is shrivelled. The main damage comes from attacks on the grain just before maturity. Although the insects leave, the damaged grain remains normal in size and remains in the wheat mixture sent to the mill.

The gluten of flour produced from infected wheat rapidly loses its cohesion upon standing in water, eventually disintegrating completely. Strong wheats resist wheat-bug attack better than soft, weak wheats do. There is little change in strong baking flours if only 1 percent of the grains are affected; in flour from soft wheats, the damage with even 1 percent to 2 percent of the grains affected can make the baking quality unacceptable. Countries in which the crop is affected by this pest include Romania, Hungary, Greece, and Morocco.

Fungus diseases. In the fungus group known as rust, the chief damage is caused by black rust (*Puccinia graminis*). Because this fungus spends part of its life on cereals and part on the barberry bush, these bushes are often eradicated near wheat fields as a preventive measure. Black rust causes cereal plants to lose their green colour and turn yellow. The grain produced is small, shrivelled, and has a low weight per bushel. New wheats, more resistant to rust, are being introduced.

In many countries wheat is attacked by smut. Stinking smut (or bunt) is fairly common in the United Kingdom. Malformed grains are produced, filled with black spores that spread over noninfected grain and give off a "fishy" smell.

Ergot (*Claviceps purpurea*) is a fungus more often attacking rye than wheat. It forms a dark purple mass, larger than the grain, containing 30 percent fatty material and the alkaloid ergotoxine, which has a profound pharmacological effect on the human and animal body and can produce abortion. Much of this fungus is likely to be removed in the mill screen room, and the clean grain sent on to the mill should contain not more than 0.04 percent of this fungus and preferably less.

HARVESTING

In the developed countries, harvesting of wheat and often other cereals is done principally by the combine harvester, though in the developing countries the ancient scythe, sickle, and flail are still widely used.

The McCormick reaper

The mechanical ancestor of today's large combines was the McCormick reaper, introduced in 1831 and followed by self-raking reapers that delivered the cut grain in bunches on the ground to be bound by hand. In 1843 a "stripper" was brought out in Australia that removed the wheat heads from the plants and threshed them in a single operation. Threshing machines were powered first by men or animals, often using treadmills, later by steam and internal-combustion engines. The modern combine harvester, originally introduced in California about 1875, came into wide use in the United States in the 1920s and '30s and in the United Kingdom in the 1940s. In 1940 the self-propelled combine was introduced. The combine cuts the standing grain, threshes out the grain from the straw and chaff, cleans the grain, and discharges it into bags or grain reservoirs. Other crops also can be worked by adaptations of the machine, and the reduction in harvesting time and labour is striking; in 1829 harvesting one acre of wheat required 14 man-hours, while the modern combine requires less than 30 minutes. In the early part of the 19th century harvesting a bushel of wheat required three man-hours' work; today it takes five minutes.

For satisfactory results, crops should not be too damp and should be reasonably ripe. If the grain contains over 14 percent moisture, as often happens in the United Kingdom and other European countries, it must be dried after harvesting under controlled conditions to avoid damage to the gluten. Rice can be combine-harvested, but because of its high moisture content (approaching 20 percent) it must be immediately dried.

GRADING

Wheat is an important commodity in international commerce, and many attempts have been made to ensure reliability in grading. In North America excellent grading allows the buyer to ascertain the type and standard of wheat acquired. Canada has statutory grades for most of its wheats. For wheat moving overseas from the terminal positions, standard export samples are used in grading.

Flour from an inferior grade is not automatically weaker than the top grade.

In the U.S. much of the wheat is officially graded, notably the hard spring and the hard winter wheats. Grading also takes place in Argentina and Australia, although it is not usually as precise as in North America. In many countries there is little commercial grading of wheat, and the buyer relies on his own testing and assessments of wheat arrivals. In Australia "fair average quality" (FAQ) indicates wheat not obviously unsatisfactory visually but takes no account of the baking strength and the character of the flour yielded. In recent years, however, considerable improvement in grading has taken place, especially when hard strong varieties are sold, as in the case of special high-protein Australian wheat from northwestern New South Wales and from Queensland.

In the U.K. there is no official wheat or barley grading as in North America. Barley is bought on appearance or by named variety. This is largely true in much of Europe, although the former Soviet Union introduced a grading system for wheat covering red spring, durum, white spring, red winter, and white winter, with special subclasses based on factors such as vitreousness, colour, and weight.

STORAGE

Cereal storage has been of concern from the earliest times; references are made to it in the Bible. Harvest variations from season to season produced carryover requiring storage, a problem that grew with increasing populations and developing commerce. The diary of Samuel Pepys (1633–1703) records the destruction of the wheat storehouses in the Great Fire of London (1666) and mentions the existence of these storehouses from the reign of Henry VIII (ruled 1509–47). With modern international cereal trade, huge silos are now found at the main points of export and at the docks of importing countries. In the major exporting countries silos at the country elevators feed the terminal silos; inefficient storage at any of these points makes the cereals highly vulnerable to insects and rodent attack. In certain regions, such as India, losses have amounted to 40 percent of the crop.

Grain respiration

A constant danger also lies in the respiration of the grain. If the moisture content of grain is low (10–12 percent), a rise in temperature resulting from respiration is unlikely; but if the bulk is large and the moisture content high (over 16 percent), the heat may not be dissipated, causing the temperature to rise and further increase the rate of respiration. Consequently, cereal stocks are turned over to ventilate the grain and to keep the temperature low. The problem also occurs in the holds of ships; much litigation has resulted from the arrival of hot and damaged cargoes.

Molds and fungi are other sources of spoilage that have received extensive study in recent years. Cleaning processes remove as much as possible of external molds before storage, but in hot countries, particularly, the problem remains serious. Under primitive conditions the habits and development of communities depended largely on their skill in storing grain.

Heat is also frequently a cause of loss of weight, loss in milling value, and loss in food value through its provision of a favourable environment for such insects as the grain weevil (*Sitophilus granarius*), the rice and maize weevils (*S. oryzae*), the lesser grain borer (*Rhizopertha dominica*), and the angoumois grain moth (*Sitotroga cerealella*). These are all endosperm borers. Among the grain germ eaters are the rust-red grain beetle (*Cryptolestes ferrugineus*), the saw-toothed grain beetle (*Oryzaephilus surina-

mensis), the khapra beetle (*Trogoderma granarium*), and the warehouse moth (*Ephestia elutella*).

Secondary pests include the mill pest known as the Mediterranean flour moth (*Anagasta kuehniella*), the confused flour beetle (*Tribolium confusum*), the rust-red flour beetle (*T. castaneum*), the flat grain beetle (*Cryptolestes pusillus*), the broad-horned flour beetle (*Gnathocerus cornutus*), the cadelle beetle (*Tenebroides mauritanicus*), and a number of miscellaneous insects, including the yellow mealworm (*Tenebrio molitor*), the Australian spider beetle, and the biscuit beetle. Of the mites that invade mills, storehouses and bakeries, the commonest is the flour mite (*Acarus siro*).

Good housekeeping, with special attention to sacks and bags and their regular cleaning and disinfecting, contributes to insect control. Frequently used insecticides include inert dusts, Pyrethrum (and synergists), gamma BHC. Other contact insecticides or fumigation may be required. The common fumigator is hydrogen cyanide, but methyl bromide and ethylene oxide have been recommended.

In Canada most of the older elevators hold 20,000 to 30,000 bushels (705 to 1,060 cubic metres) of grain, but some hold as much as 100,000 bushels (3,500 cubic metres). A Canadian elevator system at Port Cartier on the St. Lawrence River is designed for the berthing of supertankers; licensed storage capacity of this installation is 10,500,000 bushels (370,000 cubic metres). Unloading of lake vessels can be carried out at 88,000 bushels (3,100 cubic metres) an hour; the two shipping belts each have maximum capacities of 50,000 bushels (1,760 cubic metres) an hour.

In the U.S. storage facilities are similar, though the proportion of wheat exported is not as great as in Canada. Many interior terminals in the U.S. handle large amounts of grain received directly from farmers.

Storage methods in Australia have improved considerably in recent years, with increasing attention given to country storing and the modernization of terminal elevators. There has been a change from bag to bulk handling; 95 percent of the grain was bulk handled by the end of the 1960s. Huge terminal elevators operate in Sydney and Newcastle.

In Argentina large terminal elevators deal with a major export trade, but grading is not as reliable as that in North America. Argentine ports receive the wheat grown in their respective areas, which gives buyers some guidance on grade and type. Considerable quantities of corn (maize) are also exported from Argentina, with precautions taken to ensure reasonably low moisture content to prevent deterioration of cargoes in shipment.

Tilbury Grain Terminal

Handling of grain received in Europe from overseas is a large operation. The Tilbury Grain Terminal in London is a good example of modern grain handling. Capable of servicing bulk carriers of up to 65,000 tons (59,000,000 kilograms), at a maximum rate of 2,000 tons (1,800,000 kilograms) an hour, the terminal feeds adjacent mills and offers a deepwater outlet for transshipment to both rail and road. Two marine leg (dockside) elevators each have a discharge rate of 1,000 tons an hour. Normal silo capacity of 105,000 tons can be extended to 240,000 tons. The silos are 127 feet (38.7 metres) high and individual bin capacities range from 60 to 900 tons. (D.W.K.-J.)

Vegetable farming

The term vegetable in its broadest sense refers to any kind of plant life or plant product; in the narrower sense, as used in this section, however, it refers to the fresh, edible portion of a herbaceous plant consumed in either raw or cooked form. The edible portion may be a root, such as rutabaga, beet, carrot, and sweet potato; a tuber or storage stem, such as potato and taro; the stem, as in asparagus and kohlrabi; a bud, such as brussels sprouts; a bulb, such as onion and garlic; a petiole or leafstalk, such as celery and rhubarb; a leaf, such as cabbage, lettuce, parsley, spinach, and chive; an immature flower, such as cauliflower, broccoli, and artichoke; a seed, such as pea and lima bean; the immature fruit, such as eggplant, cucumber, and sweet corn (maize); or the mature fruit, such as tomato and pepper.

The popular distinction between vegetable and fruit is difficult to uphold. In general, those plants or plant parts that are usually consumed with the main course of a meal are popularly regarded as vegetables, while those mainly used as desserts are considered fruits. This distinction is applied in this section. Thus, cucumber and tomato, botanically fruits, since they are the portion of the plant containing seeds, are commonly regarded as vegetables.

This section will treat the principles and practices of vegetable farming. For a discussion of the food value and the processing of vegetables, see the article FOOD PROCESSING. For a discussion of the cultivation of other plants for human consumption, see below, *Fruit farming*.

TYPES OF PRODUCTION

Vegetable production operations range from small patches of crops, producing a few vegetables for family use or marketing, to the great, highly organized and mechanized farms common in the most technologically advanced countries.

In technologically developed countries the three main types of vegetable farming are based on production of vegetables for the fresh market, for canning, freezing, dehydration, and pickling, and to obtain seeds for planting.

Production for the fresh market. This type of vegetable farming is normally divided into home gardening, market gardening, truck farming, and vegetable forcing.

Home gardening provides vegetables exclusively for family use. About one-fourth of an acre (one-tenth of a hectare) of land is required to supply a family of six. The most suitable vegetables are those producing a large yield per unit of area. Bean, cabbage, carrot, leek, lettuce, onion, parsley, pea, pepper, radish, spinach, and tomato are desirable home garden crops.

Market gardening produces assorted vegetables for a local market. The development of good roads and of motor trucks has rapidly extended available markets; the market gardener, no longer forced to confine his operations to his local market, often is able to specialize in the production of a few, rather than an assortment, of vegetables; a transformation that provides the basis for a distinction between market and truck gardening in the mid-20th century. Truck gardens produce specific vegetables in relatively large quantities for distant markets.

Market gardening

In the method known as forcing, vegetables are produced out of their normal season of outdoor production under forcing structures that admit light and induce favourable environmental conditions for plant growth. Greenhouses, cold frames, and hotbeds are common structures used. Hydroponics, sometimes called soilless culture, allows the grower to practice automatic watering and fertilizing, thus reducing the cost of labour. To successfully compete with other fresh market producers, greenhouse vegetable growers must either produce crops when the outdoor supply is limited or produce quality products commanding premium prices.

Production for processing. Processed vegetables include canned, frozen, dehydrated, and pickled products. The cost of production per unit area of land and per ton is usually less for processing crops than for the same crops grown for market because raw material appearance is not a major quality factor in processing. This difference allows lower land value, less hand labour, and lower handling cost. Although many kinds of vegetables can be processed, there are marked varietal differences within each species in adaptability to a given method.

Specifications for vegetables for canning and freezing usually include small size, high quality, and uniformity. For many kinds of vegetables, a series of varieties having different dates of maturity is required to ensure a constant supply of raw material, thus enabling the factory to operate with an even flow of input over a long period. Acceptable processed vegetables should have a taste, odour, and appearance comparable with the fresh product, retain nutritive values, and have good storage stability. The major vegetables processed commercially are indicated in Table 3.

Vegetables raised for seed production. This type of vegetable farming requires special skills and techniques. The crop is not ready for harvest when the edible portion of

Table 3: Major Vegetables and Kinds of Processing

	canning	freezing	dehydration	pickling
Asparagus	+	−	−	−
Bean	+	+	+	−
Broccoli	−	+	−	−
Cabbage	−	−	+	+
Carrot	+	+	+	+
Celery	−	−	+	−
Cucumber	−	−	−	+
Garlic	−	−	+	−
Lima bean	+	+	−	−
Onion	−	−	+	+
Parsley	−	−	+	−
Pea	+	+	−	−
Pepper	−	−	+	+
Potato	−	+	+	−
Spinach	+	+	−	−
Sweet corn	+	+	−	−
Sweet potato	+	−	+	−
Tomato	+	−	+	−

the plant reaches the stage of maturity; it must be carried through further stages of growth. Production under isolated conditions ensures the purity of seed yield. Special techniques are applied during the stage of flowering and seed development and also in harvesting and threshing the seeds.

PRODUCTION FACTORS AND TECHNIQUES

Consumer demand factors

Profitable vegetable farming requires attention to all production operations, including insect, disease, and weed control and efficient marketing. The kind of vegetable grown is mainly determined by consumer demands, which can be defined in terms of variety, size, tenderness, flavour, freshness, and type of pack. Effective management involves the adoption of techniques resulting in a steady flow of the desired amount of produce over the whole of the natural growing season of the crop. Many vegetables can be grown throughout the year in some climates, although yield per acre for a given kind of vegetable varies according to the growing season and region where the crop is produced.

Climate. Climate involves the temperature, moisture, daylight, and wind conditions of a specific region. Climatic factors strongly affect all stages and processes of plant growth.

Temperature. Temperature requirements are based on the minimum, optimum, and maximum temperatures during both day and night throughout the period of plant growth. Requirements vary according to the type and variety of the specific crop. Based on their optimum temperature ranges, vegetables may be classed as cool-season or warm-season types. Cool-season vegetables thrive in areas where the mean daily temperature does not rise above 70° F (21° C). This group includes the artichoke, beet, broccoli, brussels sprouts, cabbage, carrot, cauliflower, celery, garlic, leek, lettuce, onion, parsley, pea, potato, radish, spinach, and turnip. Warm-season vegetables, requiring mean daily temperature of 70° F or above, are intolerant of frost. These include the bean, cucumber, eggplant, lima bean, okra, muskmelon, pepper, squash, sweet corn (maize), sweet potato, tomato, and watermelon.

Premature seeding, or bolting, is an undesirable condition that is sometimes seen in fields of cabbage, celery, lettuce, onion, and spinach. The condition occurs when the plant goes into the seeding stage before the edible portion reaches a marketable size. Bolting is attributed to either extremely low or high temperature conditions in combination with inherited traits. Specific vegetable strains or varieties may exhibit significant differences in their tendency to bolt.

Young cabbage or onion plants of relatively large size may bolt upon exposure to low temperatures near 50° to 55° F (10° to 13° C). At high temperatures of 70° to 80° F (21° to 27° C) lettuce plants do not form heads and will show premature seeding. The fruit sets of tomatoes are adversely affected by relatively low and relatively high temperatures. Tomato breeders, however, have developed several new varieties, some setting fruits at a temperature as low as 40° F (4° C) and others at a temperature as high as 90° F (32° C).

Moisture. The amount and annual distribution of rainfall in a region, especially during certain periods of development, affects local crops. Irrigation may be required to compensate for insufficient rainfall. For optimum growth and development, plants require soil that supplies water as well as nutrients dissolved in water. Root growth determines the extent of a plant's ability to absorb water and nutrients, and in dry soil root growth is greatly retarded. Extremely wet soil also retards root growth by restricting aeration. Atmospheric humidity, the moisture content of the air, also contributes moisture. Certain seacoast areas characterized by high humidity are considered especially adapted to the production of such crops as the artichoke and lima bean. High humidity, however, also creates conditions favourable for the development of certain plant diseases.

The photoperiod

Daylight. Light is the source of energy for plants. The response of plants to light is dependent upon light intensity, quality, and daily duration, or photoperiod. The seasonal variation in day length affects the growth and flowering of certain vegetable crops. Continuation of vegetative growth, rather than early flower formation, is desirable in such crops as spinach and lettuce. When planted very late in the spring, these crops tend to produce flowers and seeds during the long days of summer before they attain sufficient vegetative growth to produce maximum yields. The minimum photoperiod required for formation of bulbs in garlic and onion plants differs among varieties, and local day length is a determining factor in the selection of varieties.

Each of the climatic factors affects plant growth, and can be a limiting factor in plant development. Unless each factor is of optimum quantity or quality, plants do not achieve maximum growth. In addition to the importance of individual climatic factors, the interrelationship of all environmental factors affects growth.

Certain combinations may exert specific effects. Lettuce usually forms a seedstalk during the long days of summer, but the appearance of flowers may be delayed, or even prevented, by relatively low temperature. An unfavourable temperature combined with unfavourable moisture conditions may cause the dropping of the buds, flowers, and small fruits of the pepper, reducing the crop yield. Desirable areas for muskmelon production are characterized by low humidity combined with high temperature. In the production of seeds of many kinds of vegetables, absence of rain, or relatively light rainfall, and low humidity during ripening, harvesting, and curing of the seeds are very important.

Site. The choice of a site involves such factors as soil and climatic region. In addition, with the continued trend toward specialization and mechanization, relatively large areas are required for commercial production, and adequate water supply and transportation facilities are essential. Topography—that is, the surface of the soil and its relation to other areas—influences efficiency of operation. In modern mechanized farming, large, relatively level fields allow for lower operating costs. Power equipment may be used to modify topography, but the cost of such land renovation may be prohibitive. The amount of slope influences the type of culture possible. Fields with a moderate slope should be contoured, a process that may involve added expense for the building of terraces and diversion ditches. The direction of a slope may influence the maturation time of a crop or may result in drought, winter injury, or wind damage. A level site is generally most desirable, although a slight slope may assist drainage. Exposed sites are not suitable for vegetable farming because of the risk of damage to plants by strong winds.

Soil types

The soil stores mineral nutrients and water used by plants, as well as housing their roots. There are two general kinds of soils—mineral and the organic type called muck or peat. Mineral soils include sandy, loamy, and clayey types. Sandy and loamy soils are usually preferred for vegetable production. Soil reaction and degree of fertility can be determined by chemical analysis. The reaction of the soil determines to a great extent the availability of most plant nutrients. The degree of acid, alkaline, or neutral reaction of a soil is expressed as the pH, with a pH of 7 being neutral, points below 7 being acid, and those above 7 being alkaline. The optimum pH range for plant growth varies from one crop to another. A soil can be made more acid, or less alkaline, by applying an acid-producing chemical fertilizer such as ammonium sulfate.

The inherent fertility of soils affects production quantity, and a sound fertility program is required to maintain productivity. The ability of a soil to support plant life and produce abundant harvests is dependent on the immediately available nutrients in the soil and on the rate of release of additional nutrients that are present but not available to plants. The rate of release of these additional nutrients is affected by such factors as microbial action, soil temperature, soil moisture, and aeration. Depletion of soil fertility may occur as a result of crop removal, erosion, leaching, and volatilization, or evaporation, of nutrients.

Soil preparation and management. Soil preparation for vegetable growing involves many of the usual operations required for other crops. Good drainage is especially important for early vegetables because wet soil retards development. Sands are valuable in growing early vegetables because they are more readily drained than the heavier soils. Soil drainage accomplished by means of ditches or tiles is more desirable than the drainage obtained by planting crops on ridges because the former not only removes the excess water but also allows air to enter the soil. Air is essential to the growth of crop plants and to certain beneficial soil organisms making nutrients available to the plants.

When crops are grown in succession, soil rarely needs to be plowed more than once each year. Plowing incorporates sod, green-manure crops, and crop residues in the soil; destroys weeds and insects; and improves soil texture and aeration. Soils for vegetables should be fairly deep. A depth of six to eight inches (15 to 20 centimetres) is sufficient in most soils.

Soil management involves the exercise of human judgment in the application of available knowledge of crop production, soil conservation, and economics. Management should be directed toward producing the desired crops with a minimum of labour. Control of soil erosion, maintenance of soil organic matter, the adoption of crop rotation, and clean culture are considered important soil-management practices.

Soil erosion

Soil erosion, caused by water and wind, is a problem in many vegetable-growing regions because the topsoil is usually the richest in fertility and organic matter. Soil erosion by water can be controlled by various methods. Terracing divides the land into separate drainage areas, with each area having its own waterway above the terrace. The terrace holds the water on the land, allowing it to soak into the soil and reducing or preventing gullying. In the contouring system, crops are planted in rows at the same level across the field. Cultivation proceeds along the rows rather than up and down the hill. Strip cropping consists of growing crops in narrow strips across a slope, usually on the contour. Soil erosion by wind can be controlled by the use of windbreaks of various kinds, by keeping the soil well supplied with humus, and by growing cover crops to hold the soil when the land is not occupied by other crops.

Maintenance of the organic-matter content of the soil is essential. Organic matter is a source of plant nutrients and is valuable for its effect on certain properties of the soil. Loss of organic matter is the result of the action of micro-organisms that gradually decompose it to carbon dioxide. The addition of manures and the growing of soil-improving crops are efficient means of supplying soil organic matter. Soil-improving crops are grown solely for the purpose of preparing the soil for the growth of succeeding crops. Green-manure crops, grown especially for soil improvement, are turned under while still green and usually are grown during the same season of the year as the vegetable crops. Cover crops, raised for both soil protection and improvement, are only grown during seasons when vegetable crops do not occupy the land. When a soil-improving crop is turned under, the various nutrients that have contributed to the growth of the crop are returned to the soil, adding a quantity of organic matter. Both legumes, those plants such as peas and beans having fruits and seeds formed in pods, and nonlegumes are effective soil-improving crops. The legumes, however, are more valuable, because they contribute nitrogen as well as humus. The rate of decomposition of plant material depends on the kind of crop, its stage of growth, and soil temperature and moisture. The more succulent the material is at the time it is turned under, the more quickly it decomposes. Because dry material decomposes more slowly than green material, it is desirable to turn under soil-improving crops before they are mature, unless considerable time is to elapse between the plowing and the planting of the succeeding crop. Plant material decomposes most rapidly when the soil is warm and well supplied with moisture. If soil is dry when a soil-improving crop is turned under, little or no decomposition will occur until rain or irrigation supplies the necessary moisture.

Crop rotation

The chief benefits derived from crop rotation are the control of disease and insects and the better use of the resources of the soil. Rotation is a systematic arrangement for the growing of different crops in a more or less regular sequence on the same land. It differs from succession cropping in that rotation cropping covers a period of two, three, or more years, while in succession cropping two or more crops are grown on the same land in one year. In many regions vegetable crops are grown in rotation with other farm crops. Most vegetables grown as annual crops fit into a four-or five-year rotation plan. The system of intercropping, or companion cropping, involves the growing of two or more kinds of vegetables on the same land in the same growing season. One of the vegetables must be a small-growing and quick-maturing crop; the other must be larger and late maturing.

In the practice of clean culture, commonly followed in vegetable growing, the soil is kept free of all competing plants through frequent cultivation and the use of protective coverings, or mulches, and weed killers. In a clean vegetable field the possibility of attack by insects and disease-incitant organisms, for which plant weeds serve as hosts, is reduced.

Propagation. Propagation of crop plants, involving the formation and development of new individuals in the establishment of new plantings, is usually accomplished by the use of either seeds or the vegetative parts of plants. The first type, known as sexual propagation, is used for asparagus, bean, broccoli, cabbage, carrot, cauliflower, celery, cucumber, eggplant, leek, lettuce, lima bean, okra, onion, muskmelon, parsley, pea, pepper, pumpkin, radish, spinach, sweet corn (maize), squash, tomato, turnip, and watermelon. The second type, asexual propagation, is used for the artichoke, garlic, girasole, potato, rhubarb, and sweet potato.

Although seed cost is a small portion of the total cost of crop production, seed quality strongly affects crop success or failure. Good seed should be accurately labelled, clean, graded to size, viable, and free of diseases and insects. The reliability of the seed house is an important factor in obtaining good-quality seed. Viability, or ability to grow, and longevity, the period of viability, are characteristics of seeds of any vegetable kind. In cool, dry storage conditions, those vegetable seeds having comparatively short longevity of one to two years are okra, onion, parsley, and sweet corn. Seeds having three-year longevity are those of the asparagus, bean, carrot, leek, and pea; four-year longevity is characteristic of the beet, chard, pepper, pumpkin, and tomato seeds; longevity of five years characterizes the seeds of broccoli, cabbage, cauliflower, celery, cucumber, eggplant, lettuce, muskmelon, radish, spinach, squash, turnip, and watermelon. The dry seeds of all vegetables, when packed under vacuum in hermetically sealed cans, should remain viable for a longer period than seeds stored under less protective conditions.

Viability of seeds

Crops grown from hybrid seeds (the offspring of two or more selected parental varieties and known as F_1) yield vegetables of high quantity and quality. The hybrid-seed industry is based on the production of new seed each year from the controlled pollination of selected parents found to produce the desired combination of characters in the progeny. In the early 1980s the number of F_1 hybrids was increasing in Japan, the United States, and other technically advanced countries. The number of F_1 hybrids varied with the kind of vegetable, but none had yet been introduced for the bean, celery, lettuce, okra, parsley, or pea.

Seedbeds

Planting. Most vegetable crops are planted in the field where they are to grow to maturity. A few kinds are commonly started in a seedbed, established in the greenhouse or in the open, and transplanted as seedlings. Asparagus seeds are planted in a seedbed to produce crowns used for field setting. Some vegetables can be either directly seeded in the field or grown from transplants. These include broccoli, cabbage, cauliflower, celery, eggplant, leek, lettuce, onion, pepper, and tomato. The time and method of planting seeds and plants of a particular vegetable influence the success or failure of the crop. Important factors include the depth of planting, the rate of planting, and the spacing both between rows and between plants within a row.

Factors to be considered in determining the time of planting include soil and weather conditions, kind of crop, and desired harvest time. When more than one planting of a crop is made, the second and later plantings should be timed to provide a continuous harvest for the period desired. The soil temperature required for germination of the planted seed varies markedly with the various kinds of vegetables. Vegetables that will not germinate at a temperature below 60° F (16° C) include the bean, cucumber, eggplant, lima bean, muskmelon, okra, pepper, pumpkin, squash, and watermelon. Temperatures higher than 90° F (32° C) are not favourable for the germination of seeds of celery, lettuce, lima bean, parsley, pea, and spinach.

The quantity of seeds planted, or rate of planting, is mainly determined by the characteristics of the vegetable plant. The size of seeds affects the number of plants raised in a given area. Watermelon varieties, for example, differ in seed size expressed as weight. The Sugar Baby variety has an average weight of 1.4 ounces (41 grams) for 1,000 seeds; those of Blackstone variety average 4.4 ounces (125 grams). If the two are grown on two separate plots of the same area and 4.4 ounces of seeds of each cultivar are planted, the result would be three times as many of the Sugar Baby plants as the Blackstone type. Seed size and plant-growth pattern of a vegetable are major factors that govern the number of plants raised in a given area. The trend in the early 1980s was to increase plant population for many crops to achieve the greatest yield possible without impairing quality. As plant population increases per unit area, a point is reached at which each plant begins to compete for certain essential growth factors—*e.g.*, nutrients, moisture, and light. When the population is below the level in which competition between plants occurs, increased population will have no effect on individual plant performance, and the yield per unit area will increase in direct proportion to the increment of population. When competition for essential growth factors occurs, however, yield per plant decreases.

Transplants

Early harvest and economical use of space are the principal objectives of growing vegetable crops from transplants produced in a greenhouse or outdoor seedbed. It is easier to care for young plants of the cabbage, cauliflower, celery, onion, and tomato in small seedbeds than to sow the seeds in the place where the crop is to grow and mature. Land is free longer for another crop, and weeds, insects, diseases, and irrigation are more readily and economically controlled. The production of transplants is often a specialty of growers who sell their produce to other vegetable growers. The seeds may be planted at a rate three to six times that commonly used for a direct-seeded field. The young plants are removed for use as transplants when they reach the desired size and age, approximately 40 to 60 days after seeding.

Care of crops during growth. Practices required for a vegetable crop growing in the field include cultivation; irrigation; application of fertilizers; control of weeds, diseases, and insects; protection against frost; and the application of growth regulators if necessary.

Cultivation. Cultivation refers to stirring the soil between rows of vegetable plants. Because weed control is the most important function of cultivation, this work should be performed at the most favourable time for weed killing, when the weeds are breaking through the soil surface. When the plants are grown on ridges, it is necessary to cover the basal plant portion with soil in the case of such vegetables as asparagus, carrot, garlic, leek, onion, potato, sweet corn, and sweet potato.

Irrigation. Vegetable production requires irrigation in arid and semi-arid regions, and irrigation is frequently used as insurance against drought in more humid regions. In areas having intermittent rain for five or six months, with little or none during the remainder of the year, irrigation is essential throughout the dry season and may also be needed between rainfalls in the rainy season. The two types of land irrigation generally suited to vegetables are surface irrigation and sprinkler irrigation. A level site is required for surface irrigation, in which the water is conveyed directly over the field in open ditches at a slow, nonerosive velocity. Where water is scarce, pipelines may be used, eliminating losses caused by seepage and evaporation. The distribution of water is accomplished by various control structures, and the furrow method of surface irrigation is frequently employed because most vegetable crops are grown in rows. Sprinkler irrigation conveys water through pipes for distribution under pressure as simulated rain.

Irrigation requirements

Irrigation requirements are determined by both soil and plant factors. Soil factors include texture, structure, water-holding capacity, fertility, salinity, aeration, drainage, and temperature. Plant factors include type of vegetable, density and depth of the root system, stage of growth, drought tolerance, and plant population (see above *Irrigation and drainage*).

Fertilizer application. Soil fertility is the capacity of the soil to supply the nutrients necessary for good crop production, and fertilizing is the addition of nutrients to the soil. Chemical fertilizers may be used to supply the needed nitrogen, phosphorus, and potassium. Chemical tests of soil, plant, or both are used to determine fertilizer needs, and the rate of application is usually based on the fertility of the soil, the cropping system employed, the kind of vegetable to be grown, and the financial return that might be expected from the crop. Methods of fertilizer application include scattering and mixing with the soil before planting; application with a drill below the surface of the soil at the time of planting; row application before or at planting time; and row application during plant growth, also called side-dressing. Plowed down broadcast fertilizers have recently been used in combination with high analysis liquid fertilizers applied at planting or as a side-dressed band. Mechanical planting devices may employ fertilizer attachments to plant the fertilizer in the form of bands near the seed. For most vegetables, the bands are placed from two to three inches (five to 7.5 centimetres) from the seed, either at the same depth or slightly below the seed.

Use of herbicides

Weed control. Weeds (plants growing where they are not wanted) reduce crop yield, increase production cost, and may harbour insects and diseases that attack crop plants. Methods employed to control weeds include hand weeding, mechanical cultivation, application of chemicals acting as herbicides, and a combination of mechanical and chemical means. Herbicides, selective chemical weed killers, are absorbed by the plant and induce a toxic reaction. The amount and type of herbicide that can be safely used to protect vegetable crops depends on the tolerance of the specific crops to the chemical. Most herbicides are applied as a spray, and the appropriate time for application is determined by the composition of the herbicide and the kind of vegetable crop to be treated. Preplanting treatments are applied before the crop is planted; preemergence treatments are applied after the crop is planted but before its seedlings emerge from the soil; and postemergence treatments are applied to the growing crop at a definite stage of growth.

Disease and insect control. The production of satisfactory crops requires rigorous disease- and insect-control measures. Crop yield may be lowered by disease or insect attack, and when plants are attacked at an early stage of growth the entire crop may be lost. Reduction in the quality of vegetable crops may also be caused by diseases and insects. Grades and standards for market vegetables usually specify strict limits on the amount of disease and insect injury that may be present on vegetables in a designated grade. Vegetables remain vulnerable to insect and disease damage after harvesting, during the marketing and handling processes. When a particular plant pest is identified, the grower can select and apply appropriate control measures. Application of insect control at the times specific insects usually appear or when the first insects are noticed is usually most effective. Effective disease control usually requires preventive procedures.

Control of plant diseases

Diseases are incited by such living organisms as bacteria, fungi, and viruses. Harmful material enters the plant, develops during an incubation period, and finally causes infection, the reaction of the plant to the pathogen, or disease-producing organism. Control is possible during the inoculation and incubation phases, but when the plant reaches the infection stage it is already damaged. Typical plant diseases include mildew, leaf spots, rust, and wilt. Chemical fungicides may be used to control disease, but the use of disease-resistant plant varieties is the most effective means of control.

Vegetable breeders have developed plant varieties resistant to one or more diseases; such varieties are available for the bean, cabbage, cucumber, lettuce, muskmelon, onion, pea, pepper, potato, spinach, tomato, and watermelon.

Insects are usually controlled by the use of chemical insecticides that kill through toxic action. Many insecticides are toxic to harmful insects but do not affect bees, which are valuable for their role in pollination.

Frost protection. Frost protection may be accomplished by increasing the amount of heat radiated from the soil when frost is likely to occur. Irrigation on the day before a predicted frost provides additional moisture in the soil to increase the amount of heat given off as infrared rays. This extra heat protects thc plants from frost injury. A continuous supply of water provided by sprinkler irrigation may also protect plants from frost. As the water freezes on the plant leaves, it loses heat that is absorbed by the plant leaves, maintaining leaf temperature at 32° F (0° C). Because of the sugars and other substances in plant cells, the freezing point of cell sap is somewhat lower than 32° F.

Growth regulators. It is somctimes desirable to retard or accelerate maturity in vegetable crops. A chemical compound may be applied to prevent sprouting in onion crops. It is applied in the field sufficiently early for absorption by the still-green foliage but late enough to avoid suppressing the bulb yield. Another substance may be used to end the dormancy, or rest period, of newly harvested potato tubers intended for planting. The treated seed potatoes have uniform sprout emergence. The same substance is applied to celery from two to three weeks before harvest to elongate the stalks and increase the yield and is also used to accelerate maturity in artichokes. A chemical compound, applied when adverse weather conditions prevail during the period of fruit setting, has been used to encourage fruit set.

Harvesting. The stage of development of vegetables when harvested affects the quality of the product reaching the consumer. In some vegetables, such as the bean and pea, optimum quality is reached well in advance of full maturity and then deteriorates, although yield continues to increase. Factors determining the harvest date include the genetic constitution of the vegetable variety, the planting date, and environmental conditions during the growing season. Successive harvest dates may be obtained either by planting varieties having different maturity dates or by changing the sequence of planting dates of one particular variety. The successive method is applicable to such crops as broccoli, cabbage, cauliflower, muskmelon, onion, pea, sweet corn (maize), tomato, and watermelon. Certain varieties of the carrot, celery, cucumber, lettuce, parsley, radish, spinach, or summer squash can be sown in succession throughout most of the year in some climates, thus prolonging the harvest period. The length of time required for various vegetables to reach the harvest stage and the age of their fruit at that point is shown in Table 4.

Successive harvest dates

Hand harvesting is employed along with various mechanical aids for broccoli, cabbage, cauliflower, muskmelon, and pepper crops. Many vegetables grown for proccssing and some vegetables destined for the fresh market are mechanically harvested. Harvesting operations may be performed by a single machine in a single step for such vegetable crops as the bean, beet, carrot, lima bean, onion, pea, potato, radish, spinach, sweet corn, sweet potato, and tomato. Designers of harvesting machinery have been working to develop a multiple-picking harvester capable of adjustment for use with more than one crop. Vegetable breeders have been able to produce vegetables with characteristics suitable for machine harvesting, including compact plant growth, uniform development, and concentrated maturity.

Table 4: Market Maturity of Vegetables

	number of days from planting to markct maturity	age of fruit, in days, at market maturity
Bean	50–60	7–10
Beet	60–70	—
Broccoli*	50–80	—
Brussels sprouts*	90–100	—
Cabbage*	70–100	—
Carrot	70–80	—
Cauliflower*	60–120	—
Celery*	90–120	—
Chard	50–60	—
Chicory*	60–70	—
Cucumber	50–70	5–20
Eggplant*	75–90	20–40
Garlic	180	—
Kohlrabi*	60	—
Leek*	150	—
Lettuce*	60–80	—
Lima bean	65–80	15
Muskmelon	80–120	30–45
Okra	60	4–7
Onion*	100–150	—
Parsley	70–85	—
Pea	60–75	10–15
Pepper*	70–80	45–60
Potato	90–120	—
Pumpkin	100–120	80–100
Radish	25–50	—
Spinach	40–50	—
Summer squash	45–60	3–7
Sweet corn	75–100	20–27
Sweet potato	120–150	—
Tomato*	70–90	45–60
Turnip	45–60	—
Watermelon	85–100	40–50

*From time of transplanting.

STORAGE

Post-harvest changes

Fresh vegetables are living organisms, and there is a continuation of life processes in the vegetable after harvest. Changes that occur in the harvested, nonprocessed vegetable include water loss, conversion of starches to sugars, conversion of sugars to starches, flavour changes, colour changes, toughening, vitamin gain or loss, sprouting, rooting, softening, and decay.

Some changes result in quality deterioration; others improve quality in those vegetables that complete ripening after harvest. Postharvest changes are influenced by such factors as kind of crop, air temperature and circulation, oxygen and carbon dioxide contents and relative humidity of the atmosphere, and disease-incitant organisms. To maintain the fresh vegetable in the living state, it is usually necessary to slow the life processes, though avoiding death of the tissues, which produces gross deterioration and drastic differences in flavour, texture, and appearance.

Storage of vegetables contributes to price stabilization by carrying over produce from periods of high production to periods of low production. It also extends the period of consumption of many kinds of vegetables. Storage conditions can contribute to the preservation of the natural living state of the edible portion and to the prevention of deterioration through control of temperature, relative humidity, and the quality of the produce to be stored. Vegetables for storage must be free from mechanical, insect, and disease injury and should be at the proper stage of maturity.

Common (unrefrigerated) storage and cold (refrigerated) storage are the methods generally employed for vegetables. Common storage, lacking precise control of temperature and humidity, includes the use of insulated storage houses, outdoor cellars, or mounds. Cold storage allows precise regulation of temperature and humidity and maintenance of constant conditions by use of a refrigeration and ventilation system. Recommended storage conditions and duration for a number of different vegetables are given in Table 5. Temporary storage, suitable only for very brief storage periods, is frequently practiced in the shipping season when large lots are accumulated for carload or truck quantities. The refrigerator car or truck is a means of temporary storage used to protect produce while it is in transit. Short-term storage may last for four or six weeks. Economic factors, such as the probability that prices will increase later in the season, encourage long-term storage of such perishable vegetables as the onion, potato, and sweet potato (see Table 5).

Table 5: Recommended Storage Conditions of Vegetables

	temperature (°F)	relative humidity (percent)	approximate length of storage (days)
Asparagus	32	90–95	21–28
Bean, snap	45	85–90	8–10
Beet, topped	32	90–95	30–90
Beet, not topped	32	90–95	10–14
Broccoli	32	90–95	7–10
Brussels sprouts	32	90–95	21–28
Cabbage	32	90–95	90–120
Carrot, topped	32	90–95	120–150
Carrot, not topped	32	90–95	10–14
Casaba melon	40–50	80–85	10–42
Cauliflower	32	90–95	14–21
Celery	31–32	90–95	60–120
Cucumber	45–50	90–95	10–14
Eggplant	45–50	90–95	10
Garlic	32	70–75	180–240
Leek	32	85–90	30–90
Lettuce	32	90–95	14–21
Lima bean, shelled	32	90–95	15
Lima bean, unshelled	32	90–95	14–28
Okra	50	85–95	14
Onion	32	70–75	180–240
Pea	32	85–90	7–14
Pepper	45	85–90	8–10
Potato	38–40	85–90	180–270
Pumpkin	50–55	70–75	60–180
Spinach	32	90–95	10–14
Squash, summer	40–50	85–95	14–21
Sweet corn	31–32	90–95	4–8
Sweet potato	55–60	80–95	120–180
Tomato, ripe	40–50	85–90	7–10
Tomato, mature green	55–70	80–85	21–35
Turnip, topped	32	90–95	120–150
Watermelon	36–40	80–85	7–14

PREMARKETING OPERATIONS AND SELLING

Premarketing operations include washing, trimming, waxing, precooling, grading, prepackaging, and packaging. Vegetables often require washing after harvest to remove any adhering soil particles. Such vegetables as the beet, carrot, celery, lettuce, radish, spinach, and turnip are trimmed before washing to remove discoloured leaves or to cut back the green tops. Waxing of the cucumber, muskmelon, pepper, potato, sweet potato, and tomato gives the product a bright appearance and controls shrivelling through reduction of moisture loss.

Waxing

Precooling. Precooling, the rapid removal of heat from freshly harvested vegetables, allows the grower to harvest produce at optimum maturity with greater assurance that it will reach the consumer at maximum quality. Precooling benefits the vegetable by slowing the natural deterioration that starts shortly after harvest, slowing the growth of decay organisms and reducing wilt by retarding water loss. The major precooling methods include hydrocooling, contact icing, vacuum cooling, and air cooling. In hydrocooling the vegetable is cooled by direct contact with cold water flowing through the packed containers and absorbing heat directly from the produce. In contact icing crushed ice is placed in the package or spread over a stack of packages to precool the contents. The vacuum cooling process produces rapid evaporation of a small quantity of water, lowering the temperature of the crop to the desired level. Air cooling involves the exposure of vegetables to cold air; the air must be as cold as possible for rapid cooling but not low enough to freeze the produce exposed to the direct air blast.

The preferred method of precooling varies according to the physical characteristics of the vegetable. Hydrocooling is recommended for the asparagus, beet, broccoli, carrot, cauliflower, celery, muskmelon, pea, radish, summer squash, and sweet corn (maize); cabbage, lettuce, and spinach are suited to vacuum cooling; air cooling is preferred for bean, cucumber, eggplant, pepper, and tomato. After the produce is precooled, it is desirable to maintain low temperature by shipping in refrigerator cars or trucks, by storing in cold-storage rooms, and by refrigeration in retail stores.

Grading. Uniformity in size, shape, colour, and ripeness is of great importance in marketing any vegetable product, and can be secured through grading. The establishment of standard grades furnishes a basis of trade. Grade standards are based mainly on general appearance, size, trueness to type, and freedom from blemishes and defects.

Packaging. Prepackaging, or consumer packaging, has become a highly organized practice, often employing elaborate equipment. The product is placed in bags made of transparent film, trays or cartons overwrapped with transparent film, or mesh or paper bags. The packaging of produce in consumer packages lends itself to self-service in retail stores. The production region is often the most satisfactory location for prepackaging, especially when a packaging centre serves a large vegetable-growing area.

Master containers for consumer packages are commonly made of paperboard. Cartons, bags, baskets, boxes, crates, and hampers of various kinds and sizes are all used in packaging vegetables for marketing. The type of container is selected to fit the kind of vegetable; it furnishes a convenient means for transport, loading, and stacking, with security and economy of space. Uniform product throughout the package is an important consideration in packing vegetables.

Selling. Producers sell vegetables through various retail and wholesale practices. Retail sales are made directly to the consumer, often through roadside stands. Many growers sell most of their produce at wholesale to retail stores, to various types of buyers on local markets in nearby cities, or in regional markets. Growers located long distances from markets sell largely to wholesale dealers or jobbers.

Retail sales

Some growers have contracts with processors. Wholesale marketing arrangements are also made through auction markets in the producing regions and through cooperative organizations of producers. (W.A.W.)

Fruit farming

The subject of fruit and nut production deals with intensive culture of perennial plants, the fruits of which have economic significance (a nut is a fruit, botanically). It is one part of the broad subject of horticulture, which also encompasses vegetable growing and production of ornamentals and flowers. This section places further arbitrary limitations in that it does not encompass a number of very important perennial fruit crops covered elsewhere, including vanilla, coffee, and the oil-producing tung tree and oil palm (see FOOD PROCESSING; BEVERAGE PRODUCTION).

Botanists define a fruit in broad terms as the fleshy or dry ripened ovary surrounding the seed of a plant. A pomologist, or specialist in the science and practice of fruit growing, defines it somewhat more narrowly as the fleshy edible part of a perennial plant associated with development of the flower. A nut is any seed or fruit consisting of a kernel, usually oily, surrounded by a hard or brittle shell. Most edible nuts—*e.g.,* almond, walnut, cashew, pecan, pistachio, etc.—are well known as dessert nuts. Not all nuts are edible. Some, used as sources of oil or fat, may be regarded as oil seeds; others are used for ornament. The botanical definition of a nut, based on features of form and structure (morphology), is more restrictive: a hard, dry, one-celled, one-seeded fruit that does not split open at maturity. Among the nuts that fit both the botanical and popular conception are the acorn, chestnut, and filbert; other so-called nuts may be botanically a seed (Brazil nut), a legume (peanut [groundnut]), or a drupe (almond

Table 6: Scientific Names, Probable Original Source, and Uses of Fruits and Nuts

common name	scientific name of main crop	probable original source	principal uses*
Fruits			
Apple	*Malus pumila*	southeastern Europe, southwestern Asia	food, firewood
Apricot	*Prunus armeniaca*	China	food
Avocado (alligator pear)	*Persea americana*	Mexico to Colombia	food
Banana	*Musa sapientum* (common) *M. cavendishii* (dwarf) *M. paradisiaca* (plantain)	Asia (southeastern)	food
Blackberry	*Rubus allegheniensis*		food
Blueberry	*Vaccinium corymbosum*	North America	food
Cherry	*Prunus cerasus* (tart); *P. avium* (sweet)	western Asia and eastern Europe	food, wood
Citrus			
Orange	*Citrus sinensis*	Malay archipelago	food
Grapefruit	*C. paradisi*	Jamaica (?)	food
Lemon	*C. limon*	E. Indian archipelago	food, oil
Lime	*C. aurantifolia*	E. Indian archipelago	food, oil
Cranberry	*Vaccinium macrocarpon*	North America	food
Currant	*Ribes sativum* (red) *R. nigrum* (black)	western Europe	food
Date	*Phoenix dactylifera*	southern Iraq	food
Fig	*Ficus carica*	Mediterranean area	food
Gooseberry	*Ribes hirtellum* (American) *R. uva-crispa* (European)	North Africa, western Europe	food
Grape	*Vitis vinifera* (European) *V. labrusca* (American)	Caspian Sea area Atlantic slope, North America	food
Guava†	*Psidium guajava* (common) *P. cattleianum* (strawberry)	tropical America	food
Mango	*Mangifera indica*	East India, Burma, Assam, Malaya	food
Mangosteen†	*Garcinia mangostana*	Southeast Asia	food
Olive	*Olea europaea*	E. Mediterranean area	food, oil
Peach	*Prunus persica*	China	
Papaya	*Carica papaya*	Central America, Mexico	food, papain (meat tenderizer)
Passion fruit	*Passiflora edulis*	tropical America	food flavouring
Pineapple	*Ananas comosus*	West Indies	food
Plum	*Prunus domestica*	Caucasus area	food
Raspberry	*Rubus idaeus* (red) *R. occidentalis* (black) *R. neglectus* (purple)	eastern Asia	food
Strawberry	*Fragaria* sp. (8)	Americas	food
Nuts			
Almond (sweet)	*Prunus amygdalus*	W. temperate India and Persia	food
Almond (bitter)	*P. amygdalus,* var. *amara*	W. temperate India and Persia	flavouring extract
Almondette	*Buchanania lanzan*	India, Burma	food
Araucarian pine nut (piñon, pinyonie)	*Araucaria araucana*	Chile	food
Arnut (yer-nut, earth chestnut, hawk nut, lousy-nut)	*Bunium* species	W. Europe to Caucasus	food
Australian nut, *see* Macadamia			
Babassu nut	*Orbignya oleifera*	Brazil	food, fuel oil
Bambarra groundnut	*Voandzeia subterranea*	tropical Africa	food
Barbados nut (physic nut)	*Jatropha curcas*	tropical America	medicine
Baroba	*Diplodiscus paniculatus*	Philippines	starchy seeds boiled and eaten
Beech nut, American	*Fagus grandifolia*	E. United States	
Beech nut, European	*Fagus sylvatica*	Cent. Europe, S.W. Asia	salad oil
Ben nut	*Moringa oleifera*	India, West Indies	artists' oil, lubricant
Betel nut (areca nut, pinang)	*Areca catechu*	E. tropics	masticatory
Bladdernut	*Staphylea* species	temperate North America, S. Europe, S. Asia	necklaces
Bomah nut	*Pycnocoma macrophylla*	Africa	tanning, poison
Bonduc nut	*Caesalpinia bonduc*	Tropics	medicine, beads
Brazil nut (castanea, creamnut, para nut)	*Bertholletia excelsa*	Amazon basin, Brazil	food
Breadfruit, African	*Treculia africana*	tropical Africa	seeds ground for meal
Bread nut	*Brosimum alicastrum*	tropical America	food
Butternut (long or white walnut)	*Juglans cinerea*	E. United States, S.E. Canada	food
Butter pit, *see* Naras nut			
Candlenut	*Aleurites moluccana*	Malaysia	drying
Cashew (acajou, caja, cajou)	*Anacardium occidentale*	West Indies, Brazil	food, oil, wood, varnish, paint
Castanopsis nut (golden chinquapin, wild chestnut)	*Castanopsis* species	S.E. Asia, California	food
Chestnut	*Castanea* species	E. United States, S. Europe, N. Africa, Asia	food
Chile hazel	*Gevuina avellana*	Chile	food
Chinquapin	*Castanea* species	S.E. United States, China	food
Chufa (rush nut, earthnut, ground almond)	*Cyperus esculentus*	S. Europe	food
Cobnut, *see* Filbert			
Cobnut, Jamaican	*Omphalea diandra*	West Indies, tropical America	food, oil
Coconut	*Cocos nucifera*	Old World tropics	food, oil
Cohune nut (cahoun nut)	*Attalea cohune*	Honduras	oil
Cola nut, *see* Kola nut			
Coquilla nut	*Attalea funifera*	Brazil	turnery
Coquita nut (coker nut)	*Jubaea spectabilis*	Chile	oil, food
Coumara nut, *see* Tonka bean			
Dika nut	*Irvingia gabonensis*	W. Africa	food, oil
Doum nut (dom nut)	*Hyphaene thebaica*	N. and central Africa	turnery, vegetable ivory

Table 6: Scientific Names, Probable Original Source, and Uses of Fruits and Nuts (continued)

common name	scientific name of main crop	probable original source	principal uses*
Filbert (hazelnut)	*Corylus* species	E. North America, S.E. Europe, Balkans, W. Asia	food
Galo nut	*Anacolosa luzoniensis*	Philippines	food
Gasso nut	*Manniophyton africanum*	W. Africa, Congo	food
Ginkgo nut	*Ginkgo biloba*	China, Japan	food
Gnetum seed	*Gnetum gnemon*	tropical Asia	food
Groundnut (wild bean)	*Apios tuberosa*	North America	tubers eaten
Grugru nut (corozo nut)	*Acrocomia aculeata*	tropical South America	beads, oil
Helicia nut	*Helicia diversifolia*	Queensland, Australia	food
Hickory nut	*Carya* species	North America, China	food
Hodgsonia seed	*Hodgsonia macrocarpa*	tropical Asia	food
Hyphaene nut, *see* Doum nut			
Indian nut, *see* Pine nut			
Inoi nut	*Poga oleosa*	W. Africa	food
Jack nut	*Artocarpus heterophyllus*	India	food
Japanese walnut (heartnut, cordate walnut)	*Juglans cordiformis ailanthifolia*	Japan	food
Java almond (Luzon or Philippine nut)	*Canarium commune*	Pacific tropics	food
Jojoba nut (goat nut, sheep nut)	*Simmondsia californica*	California, Mexico	food, hair oil
Karaka nut	*Corynocarpus laevigatus*	New Zealand	food
Kola nut	*Cola acuminata, C. nitida*	W. tropical Africa	masticatory, stimulant
Kubili nut	*Cubilia blancoi*	Philippines	food
Ling (caltrop, lingko)	*Trapa bicornis*	China	food
Litchi (lychee, Chinese nut)	*Litchi chinensis*	S. China	food
Lotus seed	*Nelumbium nelumbo*	Asia	food
Lunan nut	*Otophora fruticosa*	Pacific region	food
Macadamia (Queensland nut, Australian nut)	*Macadamia* species	Australia	food
Manketti nut	*Ricinodendron rautanenii*	S. Africa	food
Marking nut	*Semecarpus anacardium*	India	ink, varnish, food
Moreton Bay chestnut (black bean)	*Castanospermum australe*	Australia	food
Naras nut (butter pit)	*Acanthosicyos horrida*	S.W. Africa	food, oil
Nicari Palm nut	*Cocos coronata*	Brazil	food
Nitta nut (nete)	*Parkia biglobosa*	tropical Africa	food
Nutmeg	*Myristica fragrans*	East Indies	spice
Olive	*Elaeocarpus ganitrus*	India	beads, ornaments
Owusa nut	*Plukenetia conophora*	W. tropical Africa	food
Oyster nut	*Telfairia pedata*	E. Africa	food
Palm chestnut	*Guilielma gasipaes*	tropical South America	food
Palm nut	*Elaeis guineensis*	W. Africa	oil
Paradise nut (sapucaia nut)	*Lecythis zabucajo*	Brazil	food
Pascualito nut (pinonchillo)	*Garcia nutans*	Mexico to Venezuela	quick-drying oil
Peanut (groundnut)	*Arachis hypogaea*	Brazil	food
Peanut, hog	*Amphicarpa monoica*	North America	food
Pecan (Illinois nut)	*Carya illinoensis*	S. United States	food
Pili nut	*Canarium ovatum*	Pacific tropics	food
Pine nut (piñon, pignolia)	*Pinus* species	S.W. United States, Europe, Asia	food
Pistachio (pistache, green almond)	*Pistacia vera*	Mediterranean basin to Iran, S.W. Asia	food
Poison nut	*Strychnos nuxvomica*	India	medicine
Quandong nut	*Fusanus acuminatus*	Australia	food
Ravensara nut (clove nutmeg)	*Ravensara aromatica*	Madagascar	spice
Rose nut (red nut)	*Hicksbeachia pinnatifolia*	Australia	ornamental, food
Sassafras nut	*Ocotea* species	South America	aromatic
Shea butter nut	*Butyrospermum parkii*	tropical Africa	food, soap oil
Singhara nut (water nut)	*Trapa bispinosa*	India, Kashmir	food
Snake nut	*Ophiocaryon paradoxum*	Guiana	charm for snakebite
Soap nut	*Sapindus saponaria*	S. Florida to N. South America	soap substitute
Soap nut, Indian	*Sapindus inocarpus*	India	soap substitute
Sterculia nut	*Sterculia foetida*	tropical Africa	food
Swarri nut (souari nut, sawarri nut, pekea nut, butter nut, piki)	*Caryocar* species	tropical America	food
Tacey nut	*Caryodendron orinocense*	Colombia	food
Taqua nut (ivory nut, vegetable ivory)	*Phytelephas macrocarpa*	Central America	ornaments, buttons
Tahiti chestnut (South Sea chestnut)	*Inocarpus edulis*	South Seas	food
Tallow nut (false sandalwood)	*Ximenia americana*	tropical Africa	food
Tallow nut, Chinese	*Sapium sebiferum*	China	wax for soap and candles
Tiger nut, *see* Chufa			
Tonka bean (tonqua or tonquin bean, coumara nut)	*Dipteryx odorata*	tropical South America	perfume
Torreya nut (kaya nut)	*Torreya nucifera*	China, Japan	food, oil
Tropical almond (myrobalan, Indian almond)	*Terminalia catappa*	S.W. Asia	food
Tung nut (wood-oil tree)	*Aleurites* species	S. China	paint, varnish, oil
Walnut, African	*Coula edulis*	W. tropical Africa, Congo	food
Walnut, Persian	*Juglans regia*	Orient	food, wood
Walnut, Black	*J. nigra*	North America	food, wood, ground shells for blasting, polishing, etc.
Water chestnut (water caltrop)	*Trapa natans*	Europe, Asia	food
Water chestnut, Chinese (matai)	*Eleocharis tuberosa*	S. China	food
Yeheb nut	*Cordeauxia edulis*	Somalia	food

*All fruits listed are eaten fresh. Most fruits can be canned, although some freeze better, such as blueberry, strawberry, raspberry, cranberry, and peach. Jellies, jams, juices, cider, and spirits are products of most fruits, although some are better adapted than others, such as apple and peach. Some fruits are adapted to drying, particularly those relatively high in sugar, such as prune, some peaches, fig, and some grapes. †Noncommercial.

and coconut). In this section the term nut is used in its broadest sense unless otherwise indicated.

This section treats the principles and practices of fruit cultivation. For a discussion of the processing of fruits, see the article FOOD PROCESSING; for information on their nutritive value, see NUTRITION.

Improvements in technology and consolidation of the fruit and nut industries in the most favoured climates of the world have been responsible for a steady increase in yield. Thus, the total acreage or number of plants devoted to various fruit and nut crops has dropped, remained about the same, or not risen in proportion to the increase in the respective crop production.

Although fruit- and nut-growing enterprises cover great ranges of climates and plant materials, their technologies have many common problems and practices. The most significant of these are discussed below.

THE VARIETY: ITS PROPAGATION AND IMPROVEMENT

Selection of plants

The first step in establishing a fruit- or nut-growing industry is the selection of individual plants with high productivity and a superior product. Such an individual is a horticultural variety. If it is multiplied vegetatively from rooted cuttings, from root pieces that throw shoots, or by graftage, each plant in the group (called a clone) that results is identical with the others. Nearly all commercially important perennial fruit and nut crops are clonally propagated; *i.e.,* their varieties are multiplied vegetatively by one means or another. Some nut crops, such as the wild pecan, cashew, black walnut, hickory, and chestnut still come from trees that grow at random from seed; hence, character and quality tend to vary.

Many important varieties of fruit plants were selected generations ago. The Sultanina (Thompson Seedless) grape, the Lob Injir (Calimyrna) fig, and the Gros Michel banana have obscure origins; planted by the millions since selection, each specimen is actually a vegetative continuation of the selected individual growing on an independent root system. But regardless of the age of a fruit-growing industry, or the perfection of some of the selected varieties, a continuing search for new varieties is essential. There is always room for improvement in climatic adaptability, in insect and disease resistance, and in the solution of special horticultural or marketing problems. In fact, government experiment stations over the world now stress scientific breeding for improvement of market quality and yield of key fruit and nut crops.

Not only are varietal selection and improvement a continuing need but so also is the maintenance of existing varieties. Although an improved vegetative mutation of a variety is exceptional, the opportunities for accidental multiplication of degenerate (low-quality) mutants increase in proportion to the number of specimens of the variety. As a result, care is taken to propagate a clone only from superior individuals, and in the case of citrus, where mutation is especially common, further precautions are necessary. There are, of course, occasional mutations that may greatly improve a variety and these are sought, selected, and propagated.

Vegetative propagation technique

Vegetative propagation technique varies with the individual fruit plant. Date, banana, and pineapple are multiplied by use of offshoots or suckers. Grape, fig, olive, currant, and blueberry are usually propagated from cuttings. Strawberry and black raspberry reproduce vegetatively by special organs—the former by stolons or runners, the latter by cane tip rooting or layering. Many kinds of fruit trees must be grafted or budded on especially grown rootstocks because the species to be multiplied does not root itself easily; apple, pear, peach, mango, and citrus are examples of this group. Many nut trees have a single taproot with but few branching roots, necessitating a deep hole and special care in transplanting.

Today's trend is toward a smaller tree in most fruit crops, particularly the apple and pear, and toward closer planting in hedgerow style, with carefully regulated fertilization and irrigation (Figure 11). This increases production per acre, lowers labour cost, increases early yields, and facilitates access in maintenance and harvesting. This approach, in fact, has been used for decades in Europe. Labour is the largest element of cost in fruit and nut production. Every means is exploited to reduce, facilitate, or eliminate hand labour.

With most fruit species a period of one to two years intervenes between the time a cutting is rooted and the time the plant is ready for setting in the field, or between graftage or budding and field planting. During this interval the plants remain in a nursery where they can be given intensive culture in rows. Pineapple and banana planting materials, however, do not require nursery care before field planting.

In choosing fruit varieties, the grower must (1) recognize the relative adaptabilities of available varieties to the climatic and soil conditions of his farm and (2) select a group that satisfies both his management needs and the market

By courtesy of (left) J & B Ltd Yakima, Wash , (top right) Tree Fruit Research Center, Washington State University, Wenatchee, (bottom right) Michigan State University, East Lansing, photographs, (top right) R Paul Larsen, (bottom right) R.F Carlson

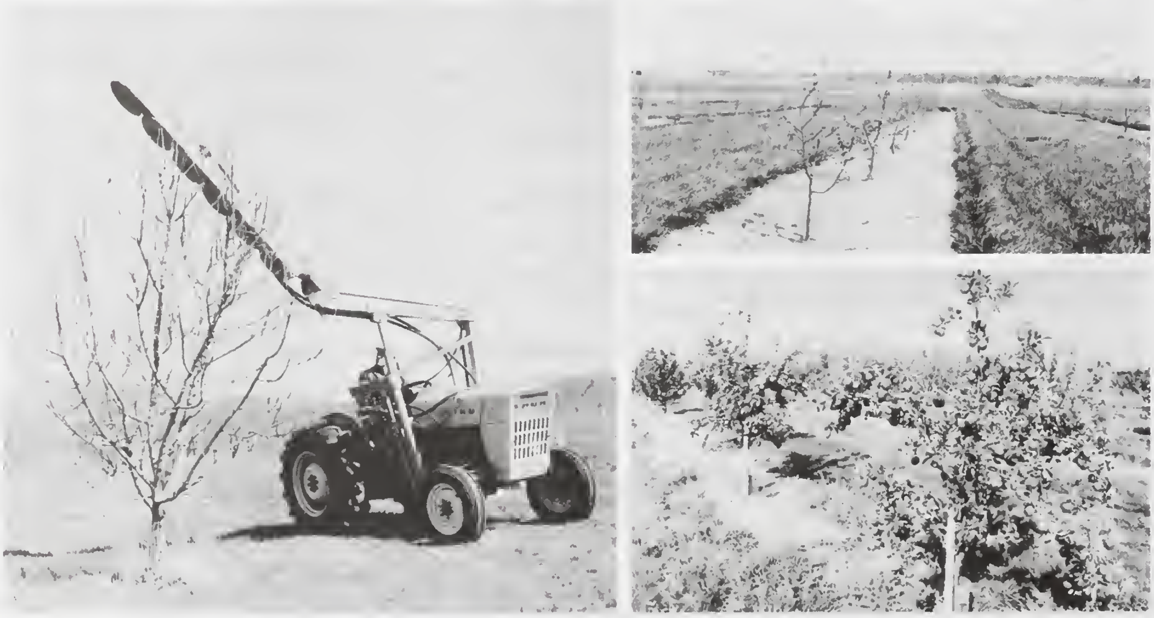


Figure 11: *Fruit growing.*
(Left) Tractor-powered fruit tree mower pruning side and top branches; further thinning is done with pneumatic shears. (Top right) A young apple orchard in which soil management involves chemical weed control under the trees and mowed sod between the tree rows. (Bottom right) Apple trees grown on dwarf stocks to produce higher yields per ground unit.

demands from those best adapted to his conditions. For instance, an apple producer in the northeastern U.S. may raise four varieties: Milton, McIntosh Red, Red Delicious, and Rome Beauty. The main harvest seasons for these succeed each other at two-week intervals; this helps him extend the harvest period and make efficient use of his labour. The first two varieties cross-fertilize satisfactorily, as do the last two. The first of these varieties is usually marketed without storage, while the storage seasons of the others are of increasing length. This helps the grower to extend his marketing period.

CULTIVATION

Site selection. The site of a fruit-growing enterprise is as significant in determining its success as the varieties grown. In fact, variety and site together set a ceiling on the productivity and profit that can be realized under the best management. In most developed fruit regions microclimatic conditions (climate at plant height, as influenced by slight differences in soil, soil covering, and elevation) and soil conditions are the two components of a site that determine its desirability for a fruit-growing enterprise. Sometimes (particularly with highly perishable fruits) transportation to market must also be considered.

Climate, soil conditions

Local conditions at a site that expose it to unusual frost hazard are as detrimental to citrus in Florida as they are to peach trees in New Zealand and apple trees in the south of England. In regions and sites where temperatures during the season may drop no more than a few degrees below freezing, artificial frost protection is sometimes used. This is accomplished by open-flame burning (petroleum bricks, logs, etc.) or heating of metal objects with oil, gas, propane, electricity, etc. (stones or stacks that radiate heat). Another technique is the spraying of water on plants (*e.g.*, strawberries) as long as the temperature is below freezing.

For highest productivity, most fruit trees must root extensively to a depth of three feet (one metre) or more. Heavy subsoil or other conditions causing imperfect internal drainage may result in shallow, weak root systems that do not take water and nutrients efficiently from the soil. In semi-arid and arid regions, accumulation of saline soils in a subsurface layer sometimes limits rooting of fruit trees, causes abnormal foliar symptoms, and reduces yields. Tiling and surface ditching help decrease water accumulation in poorly drained subsoils and reduce wet spots in otherwise satisfactory sites. Special control of irrigation procedures and periodic leaching may alleviate the worst salt effects in saline soils. Choice of tolerant species, varieties, and rootstocks may make fruit growing economical on imperfectly drained or mildly saline sites, though plants rarely perform as well as they do on sites free from these difficulties. Coconuts, however, tolerate saline soil conditions near tropical saltwater coasts.

Once selected, a site is cleared, levelled (if needed), and cultivated. Then drainage, irrigation, and road systems are installed as required. In rolling or sloping terrain, where contour planting is needed to control erosion and conserve moisture, the locations of the plant or row positions are determined by the contour terraces and waterways established. In old lands, nematode or other pest populations make fumigation necessary before planting. In some problem California soils, giant plows and treaded tractors turn the soil to depths of three to six feet (one to two metres). In very infertile sites, or sites where the physical condition of the surface soil is poor, it may be helpful to grow a succession of leguminous cover crops for a year or more before planting and/or apply a fertilizer containing major fertilizer elements (nitrogen, potassium, phosphorus, calcium, sulfur) and all or certain trace elements (iron, manganese, boron, zinc, copper, molybdenum) and lime, based on a soil test.

Planting and spacing systems. Growth, flowering habits, and light requirements on the one hand, and management problems on the other, determine the most satisfactory planting plan for a fruit- and nut-growing enterprise. There is a trend toward use of dwarfing stocks, growth control chemicals, or closer planting and training, or all of them to get the highest yields and best operation efficiency possible on a unit of ground.

Low-growing crops such as strawberry and pineapple are usually managed in beds containing several rows, or in less formal matted rows. In an acre of strawberries, 200,000 or more plants may occupy the matted rows. A pineapple plantation with two-row beds, having plants one foot (0.3 metre) apart in rows two feet (0.6 metre) apart totals 15,000 to 18,000 plants per acre (37,000 to 44,000 per hectare). With such dense populations, intense competition for light, water, and nutrients causes smaller average fruit size. Nevertheless, the total yield per unit of land is usually greater than it would be with lower plant numbers.

The spacing of grapevines along a trellis row and of trees planted in hedgerows involves the same group of problems. Maximum vineyard production frequently results with vine distances of eight to nine feet (2.4 to 2.7 metres; 600 ± per acre [1,500 per hectare]). The trend for peach trees and spur-type apple strains is hedgerows 14 feet (4.2 metres) apart or closer, in rows 18 to 20 feet (5.4 to 6 metres) apart.

With those species and varieties that require cross-pollination by insects, the planting plan must take those special needs into account. This is a problem with apple, pear, plum, and sweet cherry orchards. At least two varieties that cross-fertilize successfully must be planted in association with each other.

Training and pruning. Pruning is the removal of parts of a plant to influence growth and fruitfulness. It is an important fruit-growing practice. Primary attention is given to form in the first few years after fruit trees or vines are planted. Form influences strength and longevity of the mature plant as well as efficiency of other fruit-growing practices; pruning for form is called training. As the plant approaches maximum fruitfulness and fills its allotted space, maintenance pruning for various purposes becomes increasingly important.

The grape may be trained following one of two systems: (1) spur system, cutting growth of the previous season (canes) to short spurs, (2) long-cane system, permitting canes to remain relatively long. Whether a spur or long-cane system is followed depends on the flowering habit of the variety. Relatively small trees that respond favourably to severe annual pruning (*e.g.*, the peach and Kadota fig) are usually trained to create an open-centred tree with a scaffold of four or five main branches that originate on a short trunk and branch a number of times to provide fruiting wood. Annual renewal pruning can be reasonably efficient under these circumstances. Larger trees that do not respond favourably to heavy annual pruning are trained best to a system that encourages the main leader branch to grow erect to a height of eight to 10 feet (2.4–3.0 metres), with four or five main lateral branches at intervals on its sides forming the scaffold that carries fruiting wood up and out; this is called a modified leader system. The central leader type of tree, with one main leader up through the centre and many side branches, is common for pear and apple planted in hedgerows, and possibly for other fruits and nuts as the close-planted hedgerow system is more widely adopted.

Maintenance pruning

The principal reasons for maintenance pruning are: (1) to permit efficient spraying and harvesting operations, (2) to maintain satisfactory light exposure for most of the leaves, and (3) to create a satisfactory balance between flowering and leaf surface.

To reduce hand labour costs, larger commercial fruit growers use machine pruning (Figure 11) on many types of fruits. Peach, apple, pear, and other fruits usually planted in hedgerows are mowed across the top and sides by machine, then thinned out as needed by a follow-up crew using pneumatic clippers and hand-powered saws, operating from hydraulically manipulated scaffolds or lifts of various types.

SOIL MANAGEMENT, IRRIGATION, AND FERTILIZATION

Soil management. Two soil management practices (1) clean cultivation and chemical weed control or both and (2) permanent sod culture, illustrate contrasting purposes and effects. In clean cultivation or chemical weed control, the surface soil is stirred periodically throughout the year or a herbicide is used to kill vegetation that competes for nutrients, water, and light. Stirring increases the decomposition rate of soil organic matter and thereby releases nitrogen and other nutrients for use by the fruit crop. It may also provide some improvement in water penetration. On the other hand, laying bare the soil surface exposes it to erosion; destruction of organic matter eventually lowers fertility and causes soil structure to change from loose and friable to tight and compacted. Though sod culture minimizes the destructive processes and may permit a modest increase in fertility, the sod itself competes with fruit plants for water and nutrients and may even compete for light. As a result, permanent sod culture is practical only with tree crops that are normally rather low in vegetation, such as apple, pear, sweet cherry, nuts, and mango. Competition from established sod may be detrimental to vigorously growing fruit plants like grape, peach, and raspberry unless adequate fertilizer and water are supplied.

Because each of these soil management systems has advantages and disadvantages, modifying or complementary practices are often used; for example, cover cropping, mulching, and chemical control of vegetation with or without strip sod in the row middles. In fact, the trend is toward mowed sod middles with strip chemical control under the trees and with overhead sprinklers during hot dry weather (Figure 11). Sprinklers not only provide water but tend to cool the plants and give fruit of better market quality without aggravating diseases. Cultivation combined with winter cover cropping has been used widely in grape, peach, cherry, bush fruit, and citrus plantings, as well as with other species. Mulching is the addition of undecomposed plant materials such as straw, hay, or processors' refuse to the soil under the plants. In orchards, mulching materials are most often applied under trees maintained in permanent sod. Strip in-row chemical control of vegetation in commercial fruit plantings has almost taken over as an economical and sound practice.

Sprinklers, cultivation, mulching

Irrigation. In semi-arid and arid regions, irrigation is necessary. Probably the maximum demand occurs in date gardens, because they expose a large leaf surface the year around under conditions of high evaporation and practically no rainfall. Irrigation in humid climates is generally being provided increasingly during extended dry periods that occur at one time or another during most growing seasons. For example, large acreages of banana are irrigated on coastal lowlands of the torrid tropics where annual rainfall exceeds 60 inches (1,500 millimetres) (see above *Irrigation and drainage*).

Fertilization. Needs of perennial fruit plants for fertilizers depend on the natural fertility of the soil supporting them and on their individual requirements. Of the essential elements, supplemental nitrogen is almost always needed; potassium supplements may be needed, even in some desert areas. Although strawberry, grape, peach, and a few other fruits have responded favourably to phosphorus, and although its application has been recommended, the phosphorus requirement of woody plants is low and deficiency is rather rare. Calcium deficiency may be more common than realized; lime is often desirable to reduce soil acidity and because of other indirect benefits. Inadequate magnesium in the soil has been noted by workers studying a wide range of fruit species. Of the trace elements, zinc, iron, and boron are most likely to be deficient, but copper, manganese, and molybdenum deficiencies also are being reported for some fruits in some regions. Iron deficiency is difficult to control in orchards where soils have high alkalinity. Granulated fertilizers in modern close-planted commercial orchards are usually broadcast by machine a month or two before growth starts. Additional nitrogen sometimes is applied in heavy crop years to apple, pear, and citrus.

CROP ENHANCEMENT

Pollination. The stimulus of pollination, fertilization, and seed formation is needed to get good size, shape, and flavour of most of the fruits. (Banana, pineapple, and some citrus and fig varieties are exceptions.) Transfer of pollen from the anthers (male) to the stigmas (female) is accomplished in nature either by insects or by movement in air. It is common practice to bring beehives into the orchard during bloom. Rainy cold weather during bloom with little or no sunshine can deter activity of the honey bee (the key insect pollinator) and reduce fruit set appreciably. This is one of the main problems not fully solved by fruit researchers. Hand-pollination by daubing collected and preserved pollen onto the stigma (as is done with date palms) sometimes is practiced for other fruits, but this approach is not widespread.

Thinning. Removal of flowers or young fruit (thinning) is done to permit the remaining fruits to grow more rapidly and to prevent development of such a large crop that the plant is unable to flower and set a commercial crop the following year. Thinning is done by hand, mechanically, or chemically. With the date, the pistillate flower cluster is reduced in size at the time of hand-pollination. In the case of certain table grape varieties, some clusters are cut off. With the Thompson seedless grape, a combination of girdling the trunk bark and judicious application of gibberellin (growth regulating) sprays at blossoming gives excellent full bunches.

Young peach fruits are thinned by striking the branches with a padded pole or by shaking the entire tree for a few seconds with a well-padded motor-driven shaker arm grasping the trunk. Hand thinning of young apple and peach fruits once was also a common practice, but because of the expense and difficulty, there has been increasing use of chemical sprays as a substitute. Two kinds of sprays are used: (1) mildly caustic sprays applied during bloom, such as Elgetol in arid regions, or (2) sprays of growth-regulating substances such as 3-CPA (2,3-chlorophenoxy propionamide) applied within a few weeks after bloom in areas with late frosts.

Pest control and preservation. In many fruit enterprises, pest control is the most expensive and time-consuming growing practice. Where the concentration of fruit farms in an area warrants it, individual efforts are complemented by legislative measures including quarantine regulations to force removal of pest-laden, unattended orchards. Sometimes the most economical control procedure is biological in nature. There is increased research today to find and multiply parasites that kill fruit crop pests. Such biological methods are necessary as political pressures increase for banning DDT and other chemicals. Selection of varieties that are immune, resistant to attack, or tolerant to specific

Biological pest control

By courtesy of John Bean Division of FMC Corp., Lansing, Mich

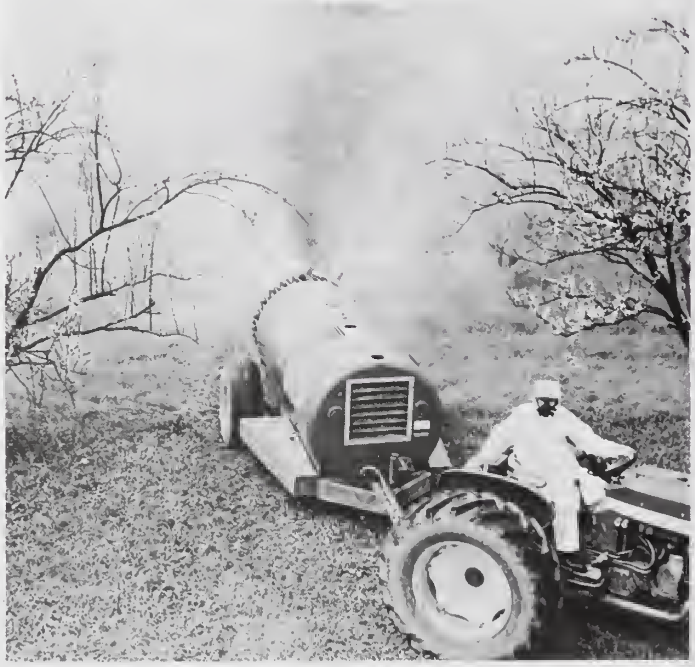


Figure 12: Air-concentrate mist blower used to spray bush fruits, grapes, and compact high-density tree fruits.

pests, is a biological control procedure also widely used. Chemical control procedures, however, are relied on most heavily. Air-blast spray or mist-application machinery covering 70 acres (28 hectares) of trees or more in a day is now in common use (Figure 12).

HARVESTING AND PACKING

The proper time to remove a fruit from the tree or plant varies with each fruit and is governed by whether the product will be sold and consumed within hours, or stored for weeks, months, or even a year. Most fruits are harvested as close as possible to the time they are eaten. A few, of which banana and pear are outstanding examples, may be harvested while immature and still ripen satisfactorily. Orange, grapefruit, and some varieties of avocado may be "stored" on the tree for several months after they have attained good quality; this method cuts costs in handling and marketing.

Many fruits, including apple, pear, orange, lemon, and grapefruit, may drop from the tree during the last part of the maturation period. Preharvest drop of these fruits can be delayed by application of dilute sprays of growth-regulating substances like naphthaleneacetic acid (NAA). The chemical spray Alar [N-(dimethylamino) succinamic acid] applied four to six weeks after bloom on apple not only reduces fruit drop at harvest but increases red colour, firmness, and return bloom the next year, in addition to other advantages.

For the fresh market, most tree and bush fruits are still harvested by hand. For processing, drying, and occasionally for fresh market, mechanical motor-driven tree and bush shakers with appropriate catching belts, bins, pallets, and electric lifts reduce harvesting and handling labour. In years to come, machinery may make it possible to machine-harvest most fruits, with no more, and possibly less, damage than with hand picking (Figure 13).

The public is becoming increasingly particular about the appearance and quality of the product it buys. Hence, store managers and suppliers seek the best grades of fruits and nuts available, and growers make every effort to produce crops with attractive colour and smooth finish. Fruits are packed by government-controlled grades such as Fancy or Extra Fancy within given size limits and are so labelled on the carton or box, together with the source. Most fruits and nuts not meeting this standard of quality are processed or sent through channels using the lower grades and off sizes.

Small packages of plastic foam or wood pulp base holding four to six fruits covered and heat sealed with polyethylene plastic film are popular. These are delivered to stores in corrugated cartons holding a few dozen packages. Citrus, apples, and whole nuts or kernels also are packaged in polyethylene bags and delivered in cartons. Loose fruit may be sold in cell cartons and tray packs consisting of stacked form-fitting pulp trays in a "bushel size" box. Every effort is made to eliminate bruising.

Shipment in controlled atmosphere

Large truck-pulled containers with individually motor-driven refrigeration units, with or without controlled atmosphere (CO_2-O_2, to retard ripening), are loaded at the fruit source and trucked to their destination or are loaded on ships by derrick for overseas shipment. These sealed containers are also being used increasingly for bananas to reduce labour and handling and to deliver the product in better condition.

Air shipment of "vine- and tree-ripe" fruit (strawberries, figs, sweet cherries, pineapples, avocados) to distances as far as from California to Europe in a day or less is becoming increasingly common with the much larger and faster cargo planes and reduced air-freight prices.

POSTHARVEST PHYSIOLOGY OF FRUITS

Fruit ripening is a form of senescence and signifies the final stage in fruit development. A fleshy fruit is the enlarged ovary of a flower (avocado) or additional floral parts such as in apple, pear, and pineapple. Usually fertilization, and sometimes pollination alone, stimulate the floral parts causing a rapid cell division that leads to differentiation and the formation of the fruit structure. During this stage fruits consist of small, young cells filled with protoplasm. When the young fruit has been stimulated, presumably by plant hormones that originate from the embryonic seeds, rapid cell expansion takes place. During this stage fruits gain rapidly in size and weight. The cells develop small cavities or spaces in their tissue (become vacuolated) and begin the process of foodstuff accumulation, which lends fruits their compositional diversity. Banana, apple, and date, for example, accumulate mainly carbohydrates. Avocado and olive store fatty materials. Important constituents of most fruits are organic acids such as malic acid, found in apple and pear; citric acid, found in citrus fruits and pineapple; and tartaric acid, found in grapes. Fruits are usually low in protein.

Stages of ripening

After cell expansion has slowed and become nominal, fruits enter the stage of maturity and undergo preparation for ripening. Some crops, such as pear and avocado, are harvested at the so-called mature-green state and allowed to ripen afterward. Most fruits are at a stage of incipient ripening before they are picked. Ripening is marked by rapid and dramatic changes that give fruits their attractive and edible character. Some of the familiar changes are softening, which results from degradation of cell wall substances; disappearance of a green background, because of chlorophyll degradation (as in pear, apple, and banana); appearance of coloured pigments such as the carotenoids—orange-yellow—and anthocyanins—red (as in orange, mango, and strawberry); a decrease in acidity and increase in the sugar content (orange, apple); and emission of the volatile substances that give many fruits

By courtesy of (left) Rutgers University, New Brunswick, N.J., (right) Chisholm Ryder Co., Inc., Niagara Falls, N.Y.

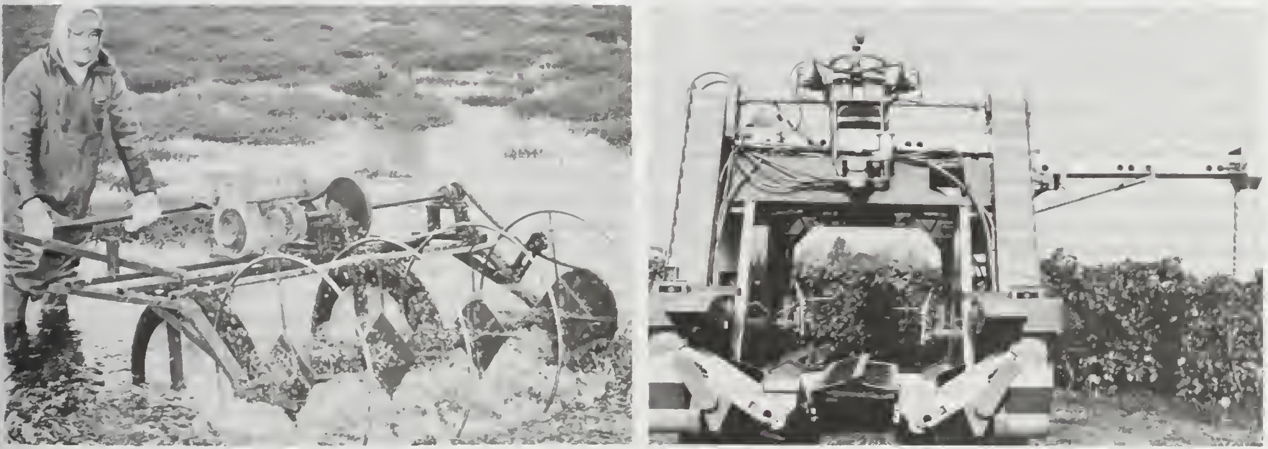


Figure 13: *Harvesting fruit mechanically.*
(Left) Cranberry harvester used on a bog to loosen and float berries. (Right) Grape harvester straddling a row of *Vitis vinifera* or *V. labrusca*. The grapes fall on belts and are carried on escalators to the top of the harvester where they enter the arm to be deposited (at right) in a tractor-pulled bulk bin for processing.

their distinct aroma (as in banana, pear, and apple). In climacteric fruits (*e.g.,* banana, pear, apple), ripening is accompanied by increased respiration. In nonclimacteric fruit (*e.g.,* strawberry, cherry) this phenomenon does not occur.

It is thought that the transition from the mature to the ripe stage is brought about by certain "ripening" enzymes. Protein molecules act as catalysts. The activity of these enzymes leads first to various ripening reactions, and then to gradual deterioration of the fruit tissue. Table 7 lists the major compositional changes in the edible fraction of several selected fruits.

Table 7: Major Compositional Changes in the Edible Fraction of Several Selected Fruits

constituent	content at maturity* (percent of fresh weight)	content upon ripening† (percent of content at maturity)
Apple		
Starch	2.0	5
Soluble sugars	7.5	99
Acids (malic)	1.0	60
Protein	0.2	120
Protopectin	0.7	12
Soluble pectin	0.2	160
Avocado		
Sugars	0.4	12
Fat	20.0	105
Protein	1.8	110
Banana		
Starch	20.0	6
Sugars	0.9	2,000
Protopectin	0.5	40
Soluble pectin	0.3	150
Orange		
Sugars	10.0	105
Acids (citric)	0.9	85
Pineapple		
Sugars	15.0	103
Acid	0.8	88

*"Content at maturity" means content when the fruit is normally harvested in mature but not necessarily ripe stage. †"Content upon ripening" refers to content at edible stage.

Because ripening leads to tissue breakdown, fruits are considered a highly perishable commodity. Different fruits have varying degrees of postharvest longevity. While strawberries last only a week to 10 days, for example, apples or lemons can be stored successfully for as long as several months.

Postharvest life of fruits can be extended by refrigeration with or without a modified oxygen–carbon dioxide atmosphere. Most temperate-zone fruits can be held safely at 32° to 41° F (0° to 5° C), but many subtropical and tropical fruits, including lemon, avocado, banana, and mango, show signs of injury from being chilled in prolonged cold storage and consequently fail to ripen properly. Bananas do not tolerate temperatures below 53° F (12° C), while several avocado varieties can be stored at temperatures as low as 46° F (8° C).

Fruit life can be extended further by both refrigeration and controlled atmosphere (CA) storage in which oxygen is kept at about 5 percent and carbon dioxide at 1 to 3 percent, while temperature is held at a level best suited to the particular fruit. So-called CA storage is common today for apples and pears and is being adapted to other fruits. Controlled atmosphere and refrigeration in conjunction with the removal of ethylene gas (which emanates from fruits and speeds ripening) helps slow the ripening process considerably. Golden Delicious apples and some pears are shipped in polyethylene containers in which a desirable, modified atmosphere is created by the respiring fruit.

Drying, canning, freezing

Drying is a standard practice for stabilizing the market movement of dates, figs, raisin grapes, prunes, and apricots. Canning is of paramount importance to the pineapple, peach, and pear industries (these fruits can be dried as well), and freezing is a means of stabilizing some of the most perishable fruits, including strawberry, raspberry, and blueberry.

Nuts are susceptible to mold, souring, staleness, discoloration, and rancidity. Cured and dried nuts are kept in prolonged cold storage under controlled temperature and humidity levels. Nuts also are stored and sold in vacuum packs of carbon dioxide-enriched atmosphere.

WASTE MATERIALS, OTHER USES

Apple wood is excellent for fireplace use, and cherry and certain other fruit woods are used for the finest household furniture. The dried residue from processing apples and citrus is made into feed for conditioning livestock for market, as are waste materials from many processed fruits. Apple pomace (waste material) is spread on the orchard floor with a manure spreader to help in soil conditioning and as a source of minerals.

Nutshells

Nutshells have many uses. Filbert shells are made into plywood, artificial wood, and linoleum; a mixture of shells with powdered coal and lignite makes cinder blocks; shells are used in making poisonous gases and gas masks, and as fuel and mulch. Cashew shell liquid, a skin irritant, is made into resins for varnishes; kills mosquito larvae; can be impregnated in wood as a varnish to preserve against insect attack; is used in automotive brake linings and clutch facings; is used as a laminating agent for paper, cloth, and glass fibres; and is used to treat cement floors and synthetic rubber to retard deterioration. Finely ground black-walnut-shell flour is used in plastic molding powder; as a glue extender; to prevent overheating of drills; to "sand"-blast jet engines; for polishing, burnishing, and deburring metal parts; for cleaning foundry molds; and to spray on tires for better traction. Pecan shells are used in place of gravel in cement walks and driveways; as fuel; as mulch and as a soil conditioner; in livestock bedding; as filler for fertilizers, feeds, etc.; in the manufacture of tanning agents, with charcoal and abrasives in hand soap; as a filler in plastic and veneer wood; and many of the same uses as black walnut shells. Some nutshells are made into beads, marbles, buttons, carving tools, ink, and ornament. The India clearing nut is cut open and rubbed on the inside of earthenware that will contain drinking water; the juice coagulates the water impurities which sink to the bottom. The nuts of the betel palm in the Far East and of the kola tree in West Africa are chewed for their stimulatory effects. (N.F.C.)

Livestock farming

Livestock is a general term that refers to all animals kept on a farm for use or for pleasure. In this section, the discussion of livestock includes both beef and dairy cattle, pigs, sheep, goats, horses, mules, asses, buffalo, and camels; the raising of birds commercially for meat or eggs (*i.e.,* chickens, turkeys, ducks, geese, guinea fowl, and squabs) is discussed below under *Poultry farming.* For further information on dairy-cattle breeds, feeding and management, see below, *Dairy farming.* For a discussion of the food value and processing of meat and dairy products, see the article FOOD PROCESSING. For a further discussion of Thoroughbred horses, see the article HORSES AND HORSEMANSHIP. Draft-horse breeds are discussed above under *Animal husbandry.*

An efficient and prosperous animal agriculture historically has been the mark of a strong, well-developed nation. Such an agriculture permits a nation to store large quantities of grains and other foodstuffs in concentrated form to be utilized to raise animals for human consumption during such emergencies as war or natural calamity. Furthermore, meat has long been known for its high nutritive value, producing stronger, healthier people.

Ruminant (cud-chewing) animals such as cattle, sheep, and goats convert large quantities of pasture forage, harvested roughage, or by-product feeds, as well as nonprotein nitrogen such as urea, into meat, milk, and wool. Ruminants are therefore extremely important; more than 60 percent of the world's farmland is in meadows and pasture. Poultry also converts feed efficiently into protein; chickens, especially, are unexcelled in meat and egg production. Milk is one of the most complete and oldest known animal foods. Cows were milked as early as 9000 BC. Hippocrates, the Greek physician, recommended milk as a medicine in the 5th century BC. Sanskrit writings from ancient India refer to milk as one of the most essential human foods.

CATTLE

Identifiable cattle breeds throughout the world number 277, with 33 generally classified as beef breeds, 18 as draft breeds, 39 as meat–draft, 54 as meat–dairy, 21 as dairy–draft, 61 as meat–dairy–draft, and 51 as dairy breeds. Most of these are quite limited in distribution and importance. No cattle are native to the North American continent, only bison, or buffalo. Cattle used for draft purposes are usually oxen—that is, castrated males at least two or three years old. Though long supplanted by the horse and the tractor in the developed countries, oxen are still used in Africa and tropical Asia.

In the U.S. today there are four times as many beef cattle as dairy cattle. Production of milk per dairy cow in the U.S. more than doubled after 1930 and reached 12,147 pounds (5,510 kilograms) in 1981. Elsewhere in the New World, the vast pampas of Argentina—an area as large as France—provide excellent conditions for grazing of cattle throughout the year without need for shelter. In Australia cattle have always ranked second in importance among livestock after sheep. Today's Australian beef cattle are raised mostly in the east and northeast portions of the continent, where feed and climate are unsuitable for sheep.

Beef cattle breeds. The British Isles led the world in the development of the principal beef breeds; Herefords, Angus, beef Shorthorns, and Galloways all originated in either England or Scotland. Other breeds of greatest prominence today originated in India (Brahman), France (Charolais; Limousin; Normandy), Switzerland (Simmental), and Africa (Africander). The Hereford breed, considered to be the first to be developed in England, probably descended from white-faced, red-bodied cattle of Holland crossed with the smaller black Celtics that were native to England and especially to Herefordshire. By the middle of the 18th century the slow process of selective breeding that resulted in the smooth, meaty, and prolific Herefords had begun. The United States statesman Henry Clay of Kentucky imported the first purebred Herefords to America in 1817.

British breeds

The Hereford, which became the most popular beef breed of the United States, is distinguished by its white face, white flanks and underline, white stockings and tail, and white crest on the neck. Its body colour ranges from cherry to mahogany red. It is of medium size, with present-day breeders making successful efforts to increase both its rate of weight gain and mature size, in keeping with the demand for cheaper, leaner beef.

The Polled Hereford is a separate breed of cattle originating from hornless mutations in 1901. It has the same general characteristics as the horned Hereford and has gained substantial favour because of its hornlessness and often faster rate of weight gain.

The Aberdeen Angus breed originated in Scotland from naturally hornless aboriginal cattle native to the counties of Aberdeen and Angus. Solid black, occasionally with a spot of white underneath the rear flanks, the breed is noted for its smoothness, freedom from waste, and high quality of meat.

Although the native home of the Galloway breed is the ancient region of Galloway in southwestern Scotland, it probably had a common origin with the Angus. The Galloway is distinguished by its coat of curly black hair. Though the breed has never attained the prominence of other beef breeds, it has been used extensively in producing blue-gray crossbred cattle, obtained by breeding white Shorthorn bulls to Galloway cows.

The beef, or Scotch, Shorthorn breed developed from early cattle of England and northern Europe, selected for heavy milk production and generally known as Durham cattle. These were later selected for the compact, beefy type by the Scottish breeders. Emphasis on leaner, high-quality carcasses in the second half of the 20th century has diminished the popularity of this breed. The Polled Shorthorn originated in 1888 from purebred, hornless mutations of the Shorthorn breed. The milking, or dual-purpose, Shorthorn, representing another segment of the parent Shorthorn breed, also was developed in England to produce an excellent flow of milk as well as an acceptable carcass, therefore resembling the original English type of Shorthorn. Shorthorns range in colour from red through roan, to white- or red-and-white-spotted.

The Brahman breed originated in India, where 30 or more separate varieties exist. Preference is given to the Guzerat, Nellore, Gir, and Krishna Valley strains, which are characterized by a pronounced hump over the shoulders and neck; excessive skin on the dewlap and underline; large, droopy ears; and horns that tend to curve upward and rearward. Their colour ranges from near white through brown and brownish red to near black. Their popularity in other areas such as South America and Europe, into which they have been imported, is attributable mainly to their heat tolerance, drought resistance, and resistance to fever ticks and other insects. The Santa Gertrudis was developed by the King Ranch of Texas by crossing Brahman and Shorthorn cattle to obtain large, hearty, tick-resistant, red cattle that have proved to be popular not only in Texas but in many regions along the semitropical Gulf Coast. Until the tick was eradicated in the southern and southwestern United States, Brahman crosses were raised almost exclusively there.

Several lesser breeds have been developed from crosses of the Brahman on other beef breeds such as: the Charbray (Charolais), Braford (Hereford), Brangus (Angus), Brahorn (Shorthorn), and Beefmaster (Brahman-Shorthorn-Hereford).

Cross-breeds

The Charolais breed, which originated in the Charolais region of France, has become quite popular in the United States for crossing on the British breeds for production of market cattle. The superior size, rate of gain, and heavy muscling of the pure French Charolais and the hybrid vigour accruing from the crossing of nonrelated breeds promise an increased popularity of this breed. Many American Charolais, however, carry significant amounts of Brahman blood, with a corresponding reduction in size, rate of gain, and muscling. Important in France, the Charolais is the foremost meat-cattle breed in Europe.

The Limousin breed, which originated in west central France, is second in importance to the Charolais as a European meat breed. Limousin cattle, often longer, finer boned, and slightly smaller than the Charolais, are also heavily muscled and relatively free from excessive deposits of fat.

The most prevalent breed of France, the Normandy, is smaller than the Charolais or Limousin and has been developed as a dual-purpose breed useful for both milk and meat production. A fourth important breed is the Maine–Anjou, which is the largest of the French breeds.

The Simmental accounts for nearly half of the cattle of Switzerland, Austria, and the western areas of Germany. Smaller than the Charolais and Limousin, the Simmental was developed for milk, meat, and draft. It is yellowish brown or red with characteristic white markings.

Beef cattle feed. Beef cattle can utilize roughages of both low and high quality, including pasture forage, hay, silage, corn (maize) fodder, straw, and grain by-products. Cattle also utilize nonprotein nitrogen in the form of urea and biuret feed supplements, which can supply from one-third to one-half of all the protein needs of beef animals. Nonprotein nitrogen is relatively cheap and abundant and is usually fed in a grain ration or in liquid supplements with molasses and phosphoric acid or is mixed with silage at ensiling time; it also may be used in supplement blocks for range cattle or as part of range pellets. Other additions to diet include corn (maize), sorghum, milo, wheat, barley, or oats. Fattening cattle are usually fed from 2.2 to 3.0 percent of their live weight per day, depending on the amount of concentrates in the ration and the rate at which they are being fattened. Such cattle gain from 2.2 to 3.0 pounds (1.0 to 1.4 kilograms) per day and require from 1.3 to 3.0 pounds (0.6 to 1.4 kilograms) of crude protein, according to their weight and stage of fattening. Up until the early 1970s, when the practice was prohibited, fattening cattle were given the synthetic hormone diethylstilbestrol as a supplement in their feed or in ear

Hormones for cattle

implants. The use of this synthetic hormone results in a 10 to 20 percent increase in daily gain with less feed required per pound of gain. Synthetic vitamin A sources have become so cheap as to permit the use of 10,000 to 30,000 International Units per day for cattle being fattened for market (finished) in enclosures bare of vegetation (drylots) used for this purpose. The economics of modern cattle finishing encourages the use of all-concentrate rations or a minimum of roughage, or roughage substitutes including oyster shells, sand, and rough plastic pellets. Corn (maize) silage produces heavy yields per acre at a low cost and makes excellent roughage for beef-cattle finishing.

Beef cows kept for the production of feeder calves are usually maintained on pasture and roughages with required amounts of protein supplement and some grain being fed only to first-calf heifers or very heavy milking cows. Most beef cows tend to be overnourished and may become excessively fat and slow to conceive unless they happen to be exceptionally heavy milkers. Most pregnant cows go into the winter in satisfactory condition and need to gain only enough to offset the weight of the fetus and related membranes. They can therefore utilize coarser roughages, having a total daily crude protein requirement of from 1.3 to 1.7 pounds (0.58 to 0.76 kilograms). Daily vitamin A supplement at the rate of 18,000 to 22,000 International Units per cow is advisable unless the roughages are of a green, leafy kind and the fall pasture has been of excellent quality. Feed requirements for bulls vary with age, condition, and activity, from 2.0 to 2.4 pounds of crude protein per day; from 25,000 to 40,000 International Units of vitamin A; and during breeding periods nearly the same energy intake as calves or short yearlings being finished for market, the main feeding requirement being to prevent their becoming excessively fat.

All cattle require salt (sodium chloride) and a palatable source of both calcium and phosphorus, such as limestone and steamed bone meal. Most commercial salts carry trace minerals as relatively cheap insurance against deficiencies that occasionally exist in scattered locations.

Beef cattle management. Beef production has become highly scientific and efficient because of the high cost of labour, land, feed, and money. Most brood-cow herds, which require a minimum of housing and equipment, are managed so as to reduce costs through pasture improvement and are typically found in relatively large areas and herds. Other aspects of management include performance testing for regular production of offspring that will gain rapidly and produce acceptable carcasses and the use of preventive medicine, feed additives, pregnancy checks, fertility testing of sires, artificial insemination of some purebred and commercial herds, protection against insects and parasites, both internal and external, adequate but not excessive feed intakes, and a minimum of handling.

Calving

Calving of beef cows is arranged to occur in the spring months to take advantage of the large supplies of cheap and high-quality pasture forages. Fall calving is less common and occurs generally in regions where winters are moderate and supplies of pasture forage are available throughout the year. Calves are normally weaned at eight to ten months of age because beef cows produce very little milk past that stage and also because they need to be rested before dropping their next calf. Feeder calves sell by the pound, so that weight for age is even more important than conformation or shape. Consequently, crossbred cattle are used; their hybrid vigour results in greater breeding efficiency and milk production on the part of the dam, as well as greater birth weight, vigour, and gaining ability on the part of the offspring.

Beef cows are normally first bred at 15 to 18 months. The gestation period is 283 days, and the interval between estrus, or periods in which the dam is in heat, is 21 days. Cows should produce a living calf every 12 months. Pasture breeding, in which nature is allowed to take its course, calls for one mature bull for every 25 cows, whereas hand breeding, in which control is exercised by the breeder, requires half as many bulls. Artificial insemination permits one outstanding sire to produce thousands of calves annually.

Diseases of beef and dairy cattle. Dairy cattle are susceptible to the same diseases as beef cattle. Many diseases and pests plague the cattle industries of the world, the more serious ones being prevalent in the humid and less developed countries. One of the more common diseases to be found in the developed countries is brucellosis, which has been controlled quite successfully through vaccination and testing. This disease produces undulant fever in humans through milk from infected cows. Leptospirosis, prevalent in warm-blooded animals and humans, is caused by a spirochete and results in fever, loss of weight, and abortion. Bovine tuberculosis has been largely eliminated; where it has not, it can infect other warm-blooded animals, including humans. Test and slaughter programs have proved effective. Rabies, caused by a specific virus that also can infect most warm-blooded animals, is usually transmitted through the bite of infected animals, either wild or domestic. Foot-and-mouth disease has been eliminated from most of North America, some Central American countries, Australia, and New Zealand. The rest of the world is still plagued by the disease, which attacks all cloven-footed animals. Humans are mildly susceptible to this organism. Successful vaccinations have been developed for blackleg, malignant edema, infectious bovine rhinotracheitis (or red nose), and several other diseases. Anaplasmosis, common to most tropical and semitropical regions, is spread by the bite of mosquitoes and flies. Anthrax, caused by a generally fatal bacterial infection, has been largely eliminated in the United States and western Europe. Rinderpest, still common to Asia and Europe, is caused by a specific virus that produces high fever and diarrhea. An infectious fever sometimes called nagana, caused by the tsetse fly, attacks both cattle and horses and is prevalent in central and southern Africa, as well as in the Philippines. Grass tetany and milk fever both result from metabolic disturbances. Bloat, caused by rapid gas formation in the rumen, or first compartment of the stomach, is sometimes fatal unless relieved. Pinkeye is an infectious inflammation of the eyes spread by flies or dust and is most serious in cattle having white pigmentation around one or both eyes. Mastitis, an inflammation of the udder, is caused by rough handling or by infection. Vibriosis, a venereal disease that causes abortion; pneumonia, an inflammation of the lungs; and shipping fever all cause serious losses and are difficult to control except through good management. Broad-spectrum antibiotics (antibiotics that are effective against various microorganisms), as well as powerful and specific pharmaceuticals, are effective and profitable means of keeping cattle herds healthy. Vermifuges, which destroy or expel parasitic worms, and insecticides, which kill harmful insects, are also highly effective and much used.

Mastitis

PIGS

Pigs are relatively easy to raise in confinement and can be slaughtered with a minimum of equipment because of their size and the many ways in which their carcasses can be processed into food and fat. Pigs are also quite efficient in converting feed to food.

Breeds. Pigs have a gestation period of 114 days with a 21-day interval between periods of estrus. A boar can mate with 15 to as many as 45 sows per year. The average litter consists of seven pigs, each with a birth weight of 2.5 pounds (1.1 kilograms). Sows should each produce 1.5 litters or more per year. Most market pigs are produced from crossbred sows by a boar of a third breed. There are more than 300 known breeds or local varieties of pigs throughout the world. A brief description follows of the better-known commercial breeds that have figured prominently in improving and upgrading domestic breeds and crosses throughout the world.

The Hampshire pig, which originated from the Norfolk thin-rind breed of England, is black with a white belt completely encircling its body and both front legs and feet. There should be no white on the head or the ham.

The Yorkshire pig, which originated early in the 19th century in England, where it was considered a bacon type, is long, lean, and trim with white hair and skin. Found in most countries, this breed is probably the most widely distributed in the world.

The Duroc-Jersey breed originated in the eastern United States from red pigs brought by Columbus and de Soto. The modern Duroc, originated from crosses of the Jersey Red of New Jersey and the Duroc of New York in the late 19th century, ranges from golden-red to mahogany-red in colour, with no black allowed. This breed proved particularly suitable for feeding in the U.S. Corn Belt (parts of Ohio, Indiana, Illinois, Wisconsin, Minnesota, South Dakota, Nebraska, Missouri, and Oklahoma; all of Iowa) and has been extensively used in Argentina, Canada, Chile, and Uruguay. The Poland China originated about 1860 in southern Ohio from a number of different breeds common to that area. The Spotted originated in Indiana about 1915 from crosses of the Poland China and the native spotted pigs. The Chester White, which originated in Chester County, Pennsylvania, after 1818, is restricted to the United States and Canada. The Berkshire, which originated in Berkshire, England, about 1770, is used for fresh pork production in England and Japan; a larger bacon type has been evolved in Australia and New Zealand. The Landrace is a white, lop-eared pig found in most countries in central and eastern Europe, with local varieties in Denmark, Germany, The Netherlands, and Sweden. World attention was first drawn to the Landrace by Denmark where, since 1890, by progeny testing (the selection of boars for breeding on the basis of the scientific assessment of their progeny), a superior pig, designed for Denmark's export trade in Wiltshire bacon to England, has been produced. Denmark no longer permits the export of Landrace for breeding. Sweden also has progeny tested from Landrace stock but for a shorter period. Pigs from Sweden were first exported to England in 1953, when prices of up to £1,000 were paid. This resulted in a worldwide Landrace explosion, and most major pig-producing countries have since taken stock.

The Landrace hog

The importance of the Asian pig breeds was recognized in the use of Chinese and "Siamese" pigs from southeastern Asia in the improvement of early European and North American breeds and is reflected in the name of the world-famous Poland China. China leads the world in pig numbers, and pork is traditional in the Chinese diet. Exports from China are substantial.

Feeding. Corn (maize) is a favourite feed for pigs, but wheat, sorghum, milo, barley, and oats also are used if the price is favourable. Wherever abundant and reasonable in price, soybean-oil meal is a favourite source of protein, and other oil meals and high-protein by-products are used in some countries. Antibiotics to control disease have become a standard ingredient in most pig rations. Improvements in breeding, disease control, management, and feed formulation have all contributed to faster gains and lower feed requirements per pound of weight gain. The use of antibiotics after World War II, especially in regions of less favourable sanitation, increased gain by as much as 20 percent.

Management. Pork production lends itself to mechanization and a minimum use of high-priced labour. The use of self-feeders and concentrated rations and construction of slotted floors and lagoons for waste disposal have become almost universal among large commerical producers in developed countries. Most commercial producers try to farrow pigs every two to three months of the year. Sows should have an 80 percent conception rate on first mating, ideally about eight or more pigs per litter, and average 1.6 litters per year. Most pigs are raised in confinement with various means of environmental control. Air-conditioned farrowing barns for excessively hot summers and heated floors and space heating or heat lamps for cold winters have become widespread.

Diseases. Pigs are subject to many infectious and parasitic diseases. Among the more common of these is transmissible gastroenteritis, an infectious disease often fatal to young pigs. Cholera, formerly controlled by vaccination, is now controlled by slaughter of infected animals. Leptospirosis, common to pigs as well as humans and most warm-blooded animals, can be controlled by vaccination. Necrotic enteritis and other infections of the intestinal tract are largely controlled by antibiotics. Atrophic rhinitis produces sneezing, crooked snouts, and unthriftiness (little or no weight gain even with proper feeding). Erysipelas, a bacterial infection of pigs, turkeys, and humans, causes in pigs inflammation of the skin, a swelling and stiffness of joints, and unthriftiness; it can be controlled through a vaccination program. Parasitic diseases are controlled mainly through effective sanitation programs and are less of a problem under confinement raising than when pigs are raised in the open. Effective vermifuges, which kill or expel parasitic worms, have been developed for further control. Production in the United States, the Scandinavian countries, and western Europe has been developed to a high state of efficiency, largely through effective control of infectious and parasitic diseases.

SHEEP

Sheep are able to subsist on sparse forage and limited water. Their wool is light in relation to its value and is relatively imperishable, both of which qualities enable wide exportation. During the 20th century, sheep-raising in some areas, particularly the western United States, has declined in favour of more profitable cattle.

Breeds. The gestation period for sheep is 147 days with 16.7 days between periods of estrus, which last 29 hours. The average number of lambs raised per hundred ewes is 91, and the average fleece weight per shearing is 8.34 pounds (3.78 kilograms).

Of more than 200 breeds of sheep in existence in the world, the majority are of limited interest except in the localities where they are raised. Sheep breeds are generally classified as medium wool, long wool, and fine wool. Of the medium wool breeds the Hampshire, Shropshire, Southdown, Suffolk, Oxford, and Dorset all originated in England. The Cheviot and Black Faced Highland originated in Scotland. The Panama, Columbia, and Targhee were developed in the United States, and the Corriedale in New Zealand. After World War II such larger breeds as the Suffolk and Hampshire increased in popularity at the expense of the smaller breeds.

The long wool breeds, including the Cotswold, Lincoln, Leicester, and Romney, were all developed in England and, in addition to mutton, produce wool of unusually long fibre length that is suitable for rugs and coarse fabrics.

Merino

The original fine-wool breed was the Merino, developed in Spain from stock native to that country before the Christian era. Though medieval Spain sought to preserve a monopoly on the Merino, the sheep gradually spread to France, Italy, and the rest of Europe. Today the Merino is prominent in Australia, the United States, Russia, South Africa, Argentina, France, and Germany; the breed is designated by various names such as Australian Merino in Australia and Merino Transhumante in Spain. The Merino was the main ancestor of the French Rambouillet, somewhat larger and less wrinkled than the Merino. This breed prospers in the western ranges of the United States, where two-thirds of that country's sheep are raised. The Corriedale breed, adapted to both farms and ranges, is especially valued in New Zealand and Australia. Most commercial sheep today represent two-breed or three-breed crosses, with white-faced crossbred ewes preferred in the range areas and a black-faced sire, such as Suffolk or Hampshire, preferred for market lambs, which are either finished for slaughter or sold as breeding ewes.

Feeding. Sheep are excellent foragers and, being ruminants, can utilize both pasture forage and harvested roughage. Selective in their grazing habits, they prefer short grass when available. Pregnant ewes can run on late pasture as long as it is available and abundant but in winter subsist satisfactorily on well-cured legume hay or mixed hay carrying a high percentage of legume. Corn (maize) silage is relatively inexpensive and relished by sheep; lactating ewes and lambs being finished for market usually require some concentrate, with corn (maize) favoured because of its high energy content and reasonable cost.

Range sheep grazing selectively on native plants frequently develop mild deficiencies of protein, energy, phosphorus, and vitamin A, especially when plants are mature or dormant or are eaten by ewes in the later stages of pregnancy or lactation. Broad spectrum antibiotics at the rate of five to 10 milligrams per pound of feed are normally used in all lamb finishing rations to prevent digestive disturbances and infections.

Management. Range sheep are normally white-faced crosses carrying both long-wool and Rambouillet breeding and are consequently very hardy and thrifty. They are wintered in bands, or flocks, of from 1,000 to 4,000 head at lower altitudes, and are moved in bands ranging from 1,000 to 1,500 head to summer range at much higher altitude, sometimes 300 miles (480 kilometres) from their winter quarters. Each flock is tended by a sheepherder and his dogs who move systematically from one grazing area to another. The herder often lives in a covered wagon or truck and may spend weeks at a time in complete solitude. The most famous sheepherders are the Basques, who emigrated widely from their home in Spain. The breeding ewes are mated to Suffolk or Hampshire rams and produce lambs during the late winter or early spring so that the lambs will be old enough to move to summer grazing without difficulty.

Shepherds

In many parts of the world small flocks are kept partly as scavengers to clean up fence rows, weeds, brush, and other undesirable forage, but this is a diminishing role. Large flocks are maintained partly for wool and partly for market lambs. Lambs are usually dropped in the spring and are sold at ages of from three to eight months and weights of around 40 pounds (18 kilograms) for Easter lambs, and 100 pounds (45 kilograms) for the usual market lambs. Sheep are sheared in the spring after the worst winter weather has passed. Some breeds are noted for producing a high percentage of twins, and others, such as the Dorset, for both high frequency of twins and heavy milk production. Shepherds frequently switch a newly born twin lamb to a ewe that has just lost a single lamb, thereby utilizing the extra milk. This practice requires skill and experience, since a mother ewe recognizes her own lamb by both its smell and the sound of its call.

Diseases. Such internal parasites as the tapeworm and several species of roundworms that infest the gastrointestinal tract are perhaps the greatest scourge of sheep, but modern vermifuges are quite effective against these. Dips are used to combat such external parasites as ticks, lice, and mites. Foot rot, caused by an infection of the soft tissue between the toes, results in extreme lameness and even loss of the hoof. The more persistent type is caused by a specific organism that is difficult to treat. The pain and the restricted movement of infected sheep result in rapid loss of weight. Enterotoxemia, or pulpy kidney, affects lambs at two to six weeks of age, especially those starting on unusually lush or rich feeds. A vaccination is quite effective in preventing this otherwise costly ailment.

GOATS

The goat has long been used as a source of milk, cheese, mohair, and meat. Its skin has been valued as a source for leather. In China, Great Britain, Europe, and North America, the goat is primarily a milk producer. By good management its limited (six months per year) breeding season and the consequent difficulty of maintaining a level supply of milk throughout the year, can be overcome. The goat is especially adapted to small-scale production of milk for the family table; one or two goats supply sufficient milk for a family throughout the year and can be maintained economically in quarters where it would not be practical to keep a cow.

Goat's milk

Pure-white goat's milk compares favourably with cow's milk in flavour and keeping qualities under sanitary conditions. It has certain characteristics differing from cow's milk that make it more easily digested by infants, invalids, and persons with allergies. Goat flesh is edible, that from young kids being quite tender and more delicate in flavour than lamb, which it resembles. Goat flesh is much prized in the Mediterranean countries, particularly in Spain, Italy, the south of France, and Greece. The Angora and Cashmere breeds are famous for their fine wool or mohair.

The many breeds may be roughly grouped: the prick-eared—*e.g.*, Swiss goats; the eastern, or Nubian, with long, drooping ears; and the wool goat—*e.g.*, Angora. While it is usually easy to distinguish goats from sheep, certain hair breeds of the latter are, to the layman, only distinguishable from goats by the direction of the tail, upward in goats, downward in sheep.

Of the Swiss goats, from which many of the best modern breeds are derived, the Toggenburg and Saanen are most important. The French breeds have much Swiss blood. In Germany the many varieties trace to Swiss breeds, which are also popular throughout Scandinavia and the Netherlands.

The Maltese goat, an important source of milk on the island of Malta, probably contains eastern blood. Many goats are found in Spain, northern Africa, and Italy, among them the Murcian, Granada, and La Mancha.

Nubians are African goats, chiefly Egyptian. They are usually large, short-haired goats with large lop ears and Roman noses. They may be of solid colour, parti-coloured, or spotted. The goats in Israel and Syria have long hair and large lop ears and most commonly are solid black or with white spots. Most Indian varieties, the best of which come from the Yamuna River area, have lop ears.

In Britain, the native goat was small, with short legs, long hair—usually gray but of no fixed colour—and with no definite markings. The widespread use of pedigree males, mostly of Swiss extraction, to improve the milk yield, has resulted in the almost total disappearance of the native types.

HORSES

Horses were among the last species of livestock to be domesticated. Domestication took place at least as early as 3000 BC, probably in the Near East. See also the article HORSES AND HORSEMANSHIP.

Breeds. The Arabian, the oldest recognized breed of horse in the world, is thought to have originated in Arabia before AD 600. Though its history is lost in the past, the breed probably descended from the Libyan horse, which, in turn, was probably preceded by horses of similar characteristics in Assyria, Greece, and Egypt as early as 1000 BC. The Arabian may be bay, gray, chestnut, brown, black, or white in hair colour, but always has a black skin. It ranges from 14.1 to 15.1 hands (4.7 to 5.0 feet, or 1.4 to 1.5 metres) in height. The Arabian horse has one lumbar vertebra less than other breeds of horse and is characterized by the high carriage of its head, long neck, and spirited action.

The Thoroughbred racing horse is descended from three desert stallions brought to England between 1689 and 1724; all of the Thoroughbreds of the world today trace their ancestry to one of these stallions.

The American Saddle Horse, which originated in the United States, was formed by crossing Thoroughbreds, Morgans, and Standardbreds on native mares possessing an easy gait. The American Saddle Horse is 15 to 16 hands (5 to 5.3 feet, or 1.5 to 1.6 metres) in height. Its colours are bay, brown, black, gray, and chestnut. There are two distinct types of the American Saddle Horse: three-gaited and five-gaited. The three natural gaits are walk, trot, and canter. Three-gaited saddle horses are shown with a short tail and cropped mane. They often have slightly less style and finish than the five-gaited horse. The five-gaited saddle horse has the three natural gaits plus two man-trained gaits—the rack and slow gait, or running walk. The American Saddle Horse is also used as a fine harness horse mainly for show.

Horse gaits

The American Quarter Horse traces to the Thoroughbred, and includes the blood of other breeds, such as the Morgan, the American Saddle Horse, and several strains of native horses. This fast, muscular horse has been raced, ridden in rodeos, and used for herding cattle.

The typical Quarter Horse is 15 to 16 hands tall and is of powerful build, suitable for both racing and the rough life of a cow pony. This horse is noted for its intelligence, easy disposition, and cow sense.

The Tennessee Walking Horse, or plantation horse, traces mainly to the Standardbred but also includes Thoroughbred and American Saddle Horse blood. The Tennessee Walking Horse is noted for its running walk, a slow-gliding gait in which the hind foot oversteps the print of the front foot by as much as 24 inches (600 millimetres). This breed is 15.2 to 16 hands high and is bay, black, chestnut, roan, or gray in colour.

The Morgan traces directly to "the Justin Morgan horse," foaled in 1793, of unknown breeding but no doubt tracing to Arabian stock. A dark bay in colour, Morgan stood 14 hands high and weighed 950 pounds (430 kilograms). He was a heavily muscled, short-legged horse of great style, quality, and endurance. He is the world's best example of prepotency, since he alone founded the Morgan breed. The Morgan is used for both riding and driving. It ranges from 14 to 16 hands in height and resembles the Arabian in size, conformation, quality, and endurance.

The American Standardbred originated around New York City during the first half of the 19th century from Thoroughbred, Morgan, Norfolk Trotter, Arabian, Barb, and pacers of mixed breeding. The modern Standardbred is smaller than the Thoroughbred, ranging from 15 to 16 hands in height and averaging about 15.2 hands. In racing condition it weighs from 900 to 1,000 pounds (410–450 kilograms). Stallions in stud condition average from 1,100 to 1,200 pounds (500–545 kilograms). Compared with the Thoroughbred, the Standardbred is longer-bodied, shorter-legged, heavier-boned, and stockier in build. Prevailing colours are bay, brown, and chestnut.

Draft horses

Draft horses have largely been supplanted by trucks and tractors in the developed countries of the world. Major draft breeds include the Percheron, developed in France; the Clydesdale of Scotland; the Shire of England; the Suffolk of England; and the Belgian of Belgium. These breeds range from 15½ to 17 hands in height at the withers; at maturity the mares weigh from 1,600 to 2,000 pounds (720–900 kilograms) and the stallions from 1,900 to 2,200 pounds (860–1000 kilograms).

The more popular pony breeds are the Shetland, which originated in the Shetland Islands, and the Hackney, of English origin. Ponies must be under 14.2 hands in height at the withers and are used both for show and for children's pleasure.

Feeding. The specific and exact nutrient requirements of horses are poorly understood. Usually, these may be supplied economically from pasture forage, harvested roughages, and concentrates. Good quality grass-legume pastures, in addition to iodized or trace-mineralized salt, will supply adequate nutrients to maintain an adult horse at light work (such as pulling a small cart) or mares during pregnancy. Lush, early spring pasture is very high in water and protein contents and may need to be supplemented with a high-energy source, such as grain, to meet the needs of horses performing medium to heavy work (such as plowing). Conversely, late fall- and winter-pasture forage is low in water and protein and may require protein and vitamin A supplementation. High-quality legume hays, such as early bloom alfalfa, are preferred for horses, especially those that are growing or lactating. Moldy or dusty feeds should be avoided because horses are extremely susceptible to forage poisoning and respiratory complications. Grass hays, such as timothy, prairie grass, orchard grass, and bluegrass, were preferred by early horsemen, especially for race horses, because they were usually free from mold and dust and tended to slow down the rate of passage through the intestinal tract. These hays are low in digestible energy and protein, however, and must be adequately supplemented. Silages of all sorts should be avoided since horses and mules are extremely susceptible to botulism and digestive upsets.

Oats are the preferred grain for horses because of their bulk. Corn (maize), barley, wheat, and milo can be used, however, whenever they are less expensive. Weanling foals require three pounds of feed per hundred pounds of live weight per day; as they approach maturity, this requirement drops to one pound of feed per hundred pounds of live weight daily. Horses normally reach mature weight at less than four years of age and 80 percent of their mature weight at less than two years of age.

A large and ever-growing number of horses stabled in cities and suburbs where sufficient roughages cannot be grown provide a large market for complete horse rations, including roughage, which are tailored to the total needs of specific animals according to their particular function at a given time, such as growth, pregnancy, lactation, or maintenance.

Horses will vary from the normal requirement in terms of weight, temperament, and previous nutrition. Foals will eat some pasture grass, forage, or hay when they are three days old and grain when they are three weeks old.

Management. Highly bred light horses are notoriously poor reproducers. Many horse farms consider a 60 percent foaling rate for a large band of mares to be average. Most large horse farms employ resident veterinarians to check for abnormalities or disease before breeding and to check mares for pregnancy 40 to 45 days after breeding. Because many mares conceive only every other year, expert assistance at foaling time is an absolute necessity, especially if the foal is sired by an expensive stallion out of a valuable mare.

Foaling mares

The gestation period of horses is 340 days. The period between estrus ranges from 18 to 28 days with an average of 22 days. Mature stallions can safely mate with from 50 to 100 mares per season, although the practice with expensive Thoroughbreds is to book no more than 35 to 40 mares.

The feet and legs of horses demand unusual attention. The old adage "no foot no horse" remains apt. Hooves should be trimmed regularly, beginning when the horse is a foal or only a few months old. Otherwise they may grow long and uneven, causing improper action, undue stress on joints, and broken or cracked hooves. Horses that are worked regularly, especially on hard and stony ground, as well as show horses and race horses in service, must be shod. Shod horses should have their hooves trimmed and their shoes refitted every four to six weeks. Tendency toward unsoundness is probably inherited but may be aggravated by poor hoof care and excessive stresses.

Diseases. Horses are especially susceptible to tetanus or lockjaw but can be given two-year protection through the use of a commonly accepted toxoid. There are two common types of abortion in horses: virus abortion, specifically viral rhinopneumonitis, and the *Salmonella* type. The former, which produces an influenza with pinkeye, catarrh, general illness, and abortion, affects both mares and foals, but all surviving horses develop natural resistance soon after infection. Pregnant mares thought to be subjected to infection may be given some protection by available vaccines. The *Salmonella* type of abortion can be prevented completely by vaccination. Encephalomyelitis, or sleeping sickness, is prevented by vaccination. A specific vaccine is available for anthrax, which is prevalent in Asia. Hemolytic anemia of foals has become a problem. Foals so afflicted are born normal but soon become sluggish and progressively weaker; the membranes of their eyes, mouth, and lips become very pale and the heartbeat becomes rapid. This condition is caused by antibodies in the mare's milk that destroy the foal's red blood cells. These antibodies are caused by the difference in blood type between the foal and the mother. Newborn foals can be muzzled to avoid nursing while their blood is checked for reaction against the serum and milk of its mother. Where reactions are noted, the mare is hand-milked at hourly intervals for 12 to 24 hours, and the foal is fed milk from another suitable mare or a milk substitute. Horses are quite susceptible to various infections, but rotation of pastures, strict sanitation, and the use of suitable vermifuges are quite effective.

ASSES AND MULES

The words donkey and ass are generally used interchangeably to denote the same animal, though ass is more often employed when the animal is wild and donkey is used for a domesticated beast. Wild asses inhabit arid, semidesert plains where the vegetation is sparse and coarse; the domestic donkey does well on coarse food and is hardy under rough conditions, hence its usefulness to man as a beast of burden in places where horses cannot flourish, such as the mountains of Ethiopia and other parts of northeast Africa, the high plains of Tibet, and the arid regions of Mongolia.

The donkey's occasional obstinacy in refusing work too heavy for it has become proverbial, but its equally proverbial stupidity has probably become legendary through its reaction to brutal treatment and neglect. It is naturally patient and persevering, responding to gentle treatment with affection and attachment to its master.

Mules are still used in some of the subtropical and tropical countries because of their ability to withstand most types of stress including heat, irregular feeding, and abuse. The mule is produced by crossing a jackass (*Equus asinus*) with a mare. The so-called Mammoth Jack was developed in America from European imports dating back to the late 18th century. It stands 15 to 16 hands (4.9 to 5.2 feet, or 1.5 to 1.6 metres) in height and weighs from 900 to 1,150 pounds (410–520 kilograms) at maturity. The reverse cross of a stallion on a jenny, or female ass, is called a hinny but theoretically has the same characteristics as a mule. At one time many different types of mules were recognized, such as draft mules, farm mules, sugar mules, cotton mules, and mining mules in declining order of size. The mining mule, a small, rugged individual weighing as little as 600 pounds (270 kilograms), was used in pit mines. These small mules are usually produced by pony dams.

Mining mules

Mules are surer-footed than horses and also more intelligent. For this reason they are still used as saddle and pack mounts in precarious terrain. Unlike horses, they also refuse to damage themselves by overeating or by thrashing around when tangled up or in cramped quarters.

BUFFALO AND CAMELS

Buffalo. The name buffalo is applied to several different cud-chewing (ruminant) mammals of the ox family (Bovidae). The true, or Indian, buffalo (*Bubalus bubalis*), also known as water buffalo, or arna, exists both as a wild and domestic animal; it has been domesticated in Asia from very early times and was introduced into Italy about the year 600. A large ox-like animal of massive and rather clumsy build with large horns that are triangular in cross section, the Indian buffalo, standing five feet (1.5 metres) at the shoulder, has a dull black body, often very sparsely covered with hair. The horns, which may be over six feet (1.8 metres) long, spread outward and upward, approaching each other toward the tips; they meet more or less in one plane above the rounded forehead and elongated face. Used for draft purposes, and also for milk and butter, the domesticated Indian buffalo is found throughout the warmer parts of the Old World from China to Egypt, and in Hungary, France, and Italy. Its cousin, the Cape, or African, buffalo (*Syncerus caffer*), a black animal of similarly massive build, has never been domesticated.

Camels. The term camel usually applies to two species of the genus *Camelus*. The Arabian camel, *Camelus dromedarius*, has one hump, the Bactrian camel, *Camelus bactrianus*, has two. The limbs are long and the feet have no traces of the second or fifth toes; the wide-spreading soft feet are well adapted for walking upon sand or snow. Horny pads on the chest and knees support the camel's weight when kneeling.

The Bactrian camel occurs throughout the highlands of Central Asia from Turkistan to Mongolia and is an important beast of burden throughout that region. The Arabian camel, characteristic of India, the Near East, and North Africa, is likewise primarily important as a beast of burden, though it also provides wool, milk, hides, and meat. It is longer-legged, shorter coated, and more lightly built than the Bactrian camel, standing about seven feet (2.1 metres) tall at the shoulder. In the 19th century camels were introduced to the U.S.–Mexico border regions, the Pacific Northwest, and Australia. The North American experiments were short-lived, but the animals were used in the exploration and development of the Australian outback until about 1940.

Camel diets

Camels can flourish on the coarsest of sparse vegetation, feeding on thorny plants, the leaves and twigs of shrubs, and dried grasses that other animals would refuse, though camels are not averse to more attractive food if it is available. When the feeding is good they accumulate stores of fat in their humps, upon which they are able to draw when conditions are adverse not only for sustenance but also for the manufacture of water by the oxidation of the fat; but they do not store water in the miscalled water cells. They are thus able to fast and go without drinking for several days; they have been known to go without water for 17 days and survive. Other adaptations that enable them to survive in deserts and other unfavourable environments include double rows of heavy protective eyelashes, haired ear openings, the ability to close their nostrils, and keen senses of sight and smell. The female produces one young at a birth after a gestation of 11 months and suckles it for a year; maturity is reached at the age of 10 to 12 years, and the life span is 30 to 40 years. (W.P.G.)

Dairy farming

Milk for human consumption is produced primarily by the cow and the water buffalo. The goat also is an important milk producer in China, India, Egypt, and in many Asian countries. Goat's milk is also produced in Europe and North America but, compared to cow's milk, goat's milk is relatively unimportant. Buffalo's milk is produced in commercial quantities in some countries, particularly India. Where it is produced, buffalo's milk is used in the same way as is cow's milk, and in some areas the community milk supply consists of a mixture of both. This section will treat the principles and practices of dairy farming. For a discussion of dairy products see the article FOOD PROCESSING.

DAIRY HERDS

Dairy cows are divided into five major breeds: Ayrshire, Brown Swiss, Guernsey, Holstein–Friesian, and Jersey. There are many minor breeds, among them the Red Dane, the Dutch Belted, and the Devon. There are also dual-purpose breeds used to produce milk and meat, notably the Milking Shorthorn and the Red Polled.

Breed characteristics

The Ayrshire breed originated in Scotland. Animals of this breed are red and white or brown and white in colour, and they are strong, vigorous, and good foragers. Ayrshire milk contains about 4.1 percent butterfat. Switzerland is the native home of the Brown Swiss. These cows are silver to dark brown in colour, with a black nose and tongue. Brown Swiss are strong and vigorous. The average fat test of the milk is 4.1 percent. The Guernsey breed originated on Guernsey Island off the coast of France. The Guernsey is fawn-coloured with clear white markings. The milk averages about 4.8 percent fat and has a deep yellow colour. The Holstein–Friesian breed originated in The Netherlands. It is black and white in colour and large in size. Holsteins give more milk than any other breed; the average butterfat content is 3.7 percent. The Jersey breed originated on the isle of Jersey, in Great Britain. Jersey cows are fawn in colour, with or without white markings. They are the smallest of the major dairy breeds, but their milk is the richest, containing on the average 5.2 percent butterfat. The protein content of milk is highest for Guernsey (3.91 percent) and Jersey (3.92 percent) and lowest for Holstein (3.23 percent).

Breeding and herd improvement. The breeds of dairy cattle have been established by years of careful selection and mating of animals to attain desired types. Increased milk and butterfat production has been the chief objective, although the objective often has shifted to increased milk and protein production. Production per cow varies with many environmental factors, but the genetic background of the cow is extremely important.

The principles of breeding to improve production have been helpful in increasing milk production in lesser developed countries. Progress has also been made in India with cows and water buffalo.

Artificial breeding has developed into a worldwide practice. Bulls with the genetic capacity to transmit high milk-producing ability to their female offspring are kept in studs. Dairy-farmer cooperatives usually operate the studs, with artificial insemination generally used. Semen for artificial insemination may be frozen for shipment to any part of the world.

Feeding dairy cattle. The dairy cow is an efficient producer of human food from roughage. This ability is attributable to a unique digestive system that consists of a four-compartment stomach capable of handling roughages not digested by human beings and other monogastric (one-stomached) animals.

Pasture is the natural feed for dairy cattle, and an abundance of good pasture provides most of the requirements of a good dairy ration. An outstanding example of grassland dairying is found in New Zealand, where cows are on pasture all year and milk production costs are at a minimum. The farmer does not need to prepare and store feed for a long winter period. Feeding a balanced ration, however, rather than grass alone, increases milk production. By 1981 the average annual production per cow in New Zealand was 7,189 pounds (3,261 kilograms) of milk, while in the U.S., where supplemental feeding is common, it was 12,147 pounds, or 5,510 kilograms. Pastures of poor quality must be supplemented with other feed, such as green crops, summer silage, or hay.

During seasons when pastures are inadequate, cows need hay, silage, and grain in sufficient amounts and balance to supply nutrient needs, and to guarantee a nutritional reserve to keep milk volume and composition from declining.

Disease prevention. Disease is one of the greatest problems of the dairy farm. It is a constant threat and may make removal of valuable animals from the herd necessary when they show even a possibility of disease. One study of removal of cows from a typical dairy herd showed that an average of 22 percent of cows were removed yearly and about a third of these were lost.

Herd management

Good herd management includes cleanliness, isolation of sick or injured animals, keeping premises free of hazards that might cause injury, and continuous protection against poisonous plants and other material. Certain diseases, such as tuberculosis, require injections. Others, such as mastitis, require constant treatment. For some diseases there is no known cure; slaughter of the animal is the only way to stop spread of the infection. Foot and mouth disease is the most notorious of these; severe measures have been employed by most governments in order to exclude or control this disease.

MILKING AND BULK HANDLING ON THE FARM

The development of milk-producing tissue in the mammae is triggered by conception; minimal production begins in the seventh or eighth week, but secretion is inhibited until after calving. The stimulus of calving increases lactation for several weeks, until another conception prompts a gradual decline. In response to pregnancy hormones and the needs of the fetus, the animal is usually dry for the month or two preceding calving. Milk is produced by the cow from her blood, and a large amount of food is necessary for maintenance of a high producing cow. The products of digestion and absorption enter the blood and are carried to the udder. There the raw materials are collected and changed into milk components. Each time the blood passes through the udder a small fraction of the components is removed to make the milk. Some 400 pounds (50 gallons, or about 200 litres) of blood must pass through the udder to make one pound (about 0.45 kilograms) of milk. A daily flow through the udder of 10 tons (20,000 pounds, or about 9,000 kilograms) of blood is required for a cow producing 50 pounds (22.5 kilograms) of milk per day. The energy required to produce milk components and to circulate the blood indicates the great importance of proper and abundant feed.

Use of milking machines

Today, most milking is done with machines by a carefully trained operator, usually twice a day, in stanchion barns or milking parlours. An experienced milker handles one to three machine units. The cows are first cleaned, and the teat cups put on. A pulsating vacuum draws the milk into a receiver or through piping into the farm milk tank (see Figure 14).

Milk is an extremely perishable commodity that must be cooled to 50° F (10° C) or less within two hours and must be maintained at that temperature until it is delivered to the consumer.

Milk is transported from farm to plant in a variety of ways, depending on the part of the world. In the Gujarat region of India, the milk is carried to a receiving station in jars on the heads of women who do the milking. The receiving station transports the milk in large cans to the plant by truck.

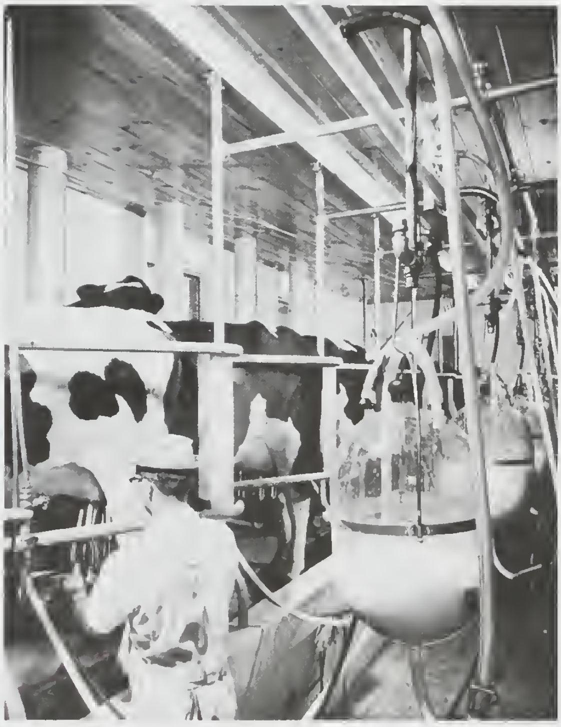


Figure 14: Attaching automatic milkers in a modern milking parlour in the United States.

Grant Heilman

In the major milk-producing countries of the world the milk is held cold in the farm tank or in cans until it is picked up, usually once or twice daily, by tanker or truck. Tankers pump the milk in at the farm and out into plant tanks on delivery. The tanker driver measures and samples each farmer's milk; fat and bacteria tests are run at the plant. The use of pipelines has been introduced on a small scale in some European countries for delivery of milk from farm to factory. (B.H.We./Ed.)

Poultry farming

This section will treat the principles and practices of poultry farming. For a discussion of the food value and processing of poultry products, see the article FOOD PROCESSING.

CHICKENS

Man first domesticated chickens of Indian origin for the purpose of cockfighting in Asia, Africa, and Europe. Very little formal attention was given to egg or meat production. Cockfighting was outlawed in England in 1849 and in most other countries thereafter. Exotic breeds and new standard breeds of chickens proliferated in the years to follow, and poultry shows became very popular. From 1890 to 1920 chicken raisers stressed egg and meat production, and commercial hatcheries became important after 1920.

Breeds. The breeds of chickens are generally classified as American, Mediterranean, English, and Asiatic. The American breeds of importance today are the Plymouth Rock, the Wyandotte, the Rhode Island Red, and the New Hampshire. The Barred Plymouth Rock, developed in 1865 by crossing the Dominique with the Black Cochin, has grayish-white plumage crossed with dark bars. It has good size and meat quality and is a good layer. The White Plymouth Rock, a variety of the Barred Plymouth Rock, has white plumage and is raised for its meat. Both varieties lay brown eggs. The Wyandotte, developed in 1870 from five or more strains and breeds, has eight varieties and is characterized by a plump body, excellent meat, and good egg production. Only the white strain is of any significance today because it is used in broiler crosses where its white plumage, quality of flesh, and rapid growth are highly desirable.

An American breed, the Rhode Island Red, developed in 1857 from Red Malay game fowl crossed with reddish-coloured Shanghais—with some brown Leghorn, Cornish, Wyandotte, and Brahma blood—is good for meat production and is one of the top meat breeds for the production of eggs. It has brilliant red feathers and lays brown eggs.

The New Hampshire, developed in the U.S. in 1930 from Rhode Island Red stock, is a meaty, early maturing breed with light-red feathers and lays large brown eggs. The only Mediterranean breed of importance today is the Leghorn. This breed, originated in Italy, has 12 varieties, the single-comb White Leghorn being more popular than all of the other types combined. This breed, the leading egg producer of the world, lays white eggs and is kept in large numbers in England, Canada, Australia, and the U.S. The White Minorca, a second Mediterranean breed, is often used in crossbreeding for egg production.

The only English breed of modern significance is the Cornish, a compact and heavily meated bird used in crossbreeding programs for broiler production. It is a poor producer of eggs, however.

The only Asiatic breed of significance today, the Brahma, which originated in India, has three varieties, the light Brahma being preferred because of its size.

Chicken breeding is an outstanding example of the application of basic genetic principles of inbreeding, linebreeding, and crossbreeding, as well as of intensive mass selection to effect faster and cheaper gains in broilers and maximum egg production for the egg-laying strains. Maximum use of heterosis, or hybrid vigour, through incrosses and crossbreeding has been made. Crossbreeding for egg production has used the single-comb White Leghorn, the Rhode Island Red, the New Hampshire, the Barred Plymouth Rock, the White Plymouth Rock, the Black Australorp, and the White Minorca. Crossbreeding for broiler production has used the White Plymouth Rock or New Hampshire crossed with White or Silver Cornish or incrosses utilizing widely diverse inbred strains within a single breed. Rapid and efficient weight gains, and high quality, plump, meaty carcasses have been achieved thereby.

Development of the egg

The male sperm lives in the hen's oviduct for two to three weeks. Eggs are fertilized within 24 hours after mating. Yolks originate in the ovary and grow to about 1.6 inches (4.0 centimetres) in diameter, after which they are released into the oviduct, where the thick white and two shell membranes are added. The egg then moves into the uterus where the thin white and the shell are added. This process requires a total of 24 hours per egg. The hatching of fertilized eggs requires 21 days, with the heavy breeds requiring a few more hours and the lighter breeds slightly fewer. Ideal hatching temperature approximates 100° F (38° C) with control of air flow, humidity, oxygen, and carbon dioxide being essential. Standardized egg-laying tests and official random sample tests have been used for many years to measure actual productivity.

Feeding. Chicken feeding is a highly perfected science that ensures a maximum intake of energy for growth and fat production. High quality and well-balanced protein sources produce a maximum amount of muscle, organ, skin, and feather growth. The essential minerals produce bones and eggs; 3 to 4 percent of the live bird being composed of minerals and 10 percent of the egg. Calcium, phosphorus, sodium, chlorine, potassium, sulfur, manganese, iron, copper, cobalt, magnesium, and zinc are all required. Vitamins A, C, D, E and K and all 12 of the B vitamins are also required. Water is essential, and antibiotics are almost universally used to stimulate appetite, control harmful bacteria, and prevent disease. Modern rations produce a pound of broiler on about two pounds (0.9 kilograms) of feed and a dozen eggs from 4½ pounds (2.0 kilograms) of feed.

Management. Among the world's agricultural industries, meat chicken breeding in the U.S. is one of the most advanced. It is presently considered the model for other animal industries, the broiler industry leading the way in advanced agricultural technology and efficiency. Intensive nutritional research and application, highly improved breeding stock, intelligent management, and scientific disease control have gone into the effort to give a modern broiler of uniformly high quality produced at ever-lower cost. Today, one person can care for 25,000 to 50,000 broilers that reach market weight in three months' time, giving an annual output of from 100,000 to 200,000 broilers. A modern broiler chick gains over 43 times its initial weight in an eight-week period. Aggressive marketing methods increased the per capita consumption of broilers more than fivefold in the three decades beginning in 1950, with further substantial increases predicted for the future. Less than half as much feed is now required to produce a pound of broiler meat as was needed in 1940. While per capita consumption of eggs has declined, the feed requirement per dozen eggs is only slightly more than half as high as it was in the early 1900s. Annual egg production per hen has increased from 104 to 244 since 1910.

Cages for laying

A carefully controlled environment that avoids crowding, chilling, overheating, or frightening is almost universal in chicken raising. Cannibalism, which expresses itself as toe picking, feather picking, and tail picking, is controlled by debeaking at one day of age and by other management practices. The feeding, watering, egg gathering, and cleaning operations are highly mechanized. More than 90 percent of the 4,200,000,000 chicks hatched per year in the early 1980s were used for broiler production and the remainder for egg production. In egg production feed represents more than two-thirds of the cost. Pullet (immature hen) flocks predominate. Hens are usually housed in wire cages with two or three hens per cage and three or four tiers of cages superposed to save space. Cages for laying hens have been found to increase production, lower mortality, reduce cannibalism, lower feeding requirements, reduce diseases and parasites, improve culling, and reduce both space and labour requirements.

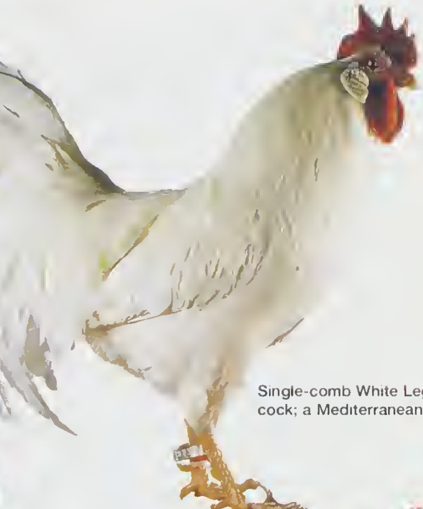
OTHER POULTRY

These include turkeys, ducks, geese, guinea fowl, and squabs.

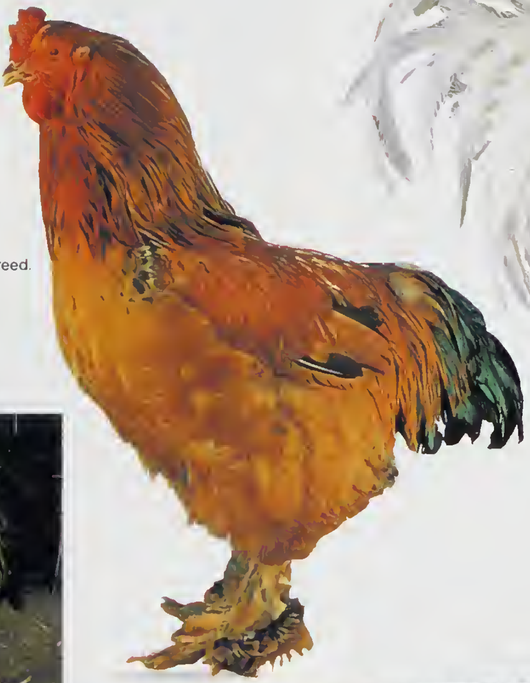
Turkey production. After World War II turkey production became highly specialized, with larger flocks predominating. Turkeys are raised in great numbers in Canada where their ancestors still live wild, as also in some parts of the U.S. Broad Breasted Bronze, Broad Breasted White, and White Holland are the most popular of the larger breeds, representing nearly three-fourths of the total production. The Beltsville Small White is the most popular of the smaller breeds and composes the bulk of the remaining 25 percent. At 24 weeks of age the toms are 50 percent heavier than the hens. In breeding flocks, one tom is required per eight or 10 hens. Tremendous improvements both in breeding and nutrition have been made in this century. Since 1910, the amount of feed required to produce a pound of turkey meat has fallen 40 percent, while the time required has been reduced 25 percent. Fifty to 80 pounds (23–36 kilograms) of feed will produce a turkey for market weight with from 2½ to 3 pounds required per pound of gain on full-size turkeys, and 2½ to 2¾ pounds (1.1–1.2 kilograms) of feed per pound (0.45 kilograms) of gain for turkey broilers, which are marketed at from 12 to 15 weeks of age. Turkey poults are hard to start on feed. One method is to dip their beaks in water and then in feed. Another is to light the feed troughs very brightly and to use oatmeal or ground yellow corn sprinkled on top of the feed. Turkeys are given range, or open land, and automatic waterers, self-feeders, range shelters, heavy fencing, and rotated pastures are used. Successful marketing techniques have increased turkey consumption; *e.g.*, in the U.S., per capita consumption from 1930/34 to 1980 rose 500 percent.

Duck and goose production. Duck raising is practiced on a limited scale in nearly all countries, for the most part as a small-farm enterprise. The flocks once kept in England are much reduced, the demand for eggs being greatly lessened, though a limited market still exists. Khaki Campbell and Indian Runner ducks are prolific layers, each averaging 300 eggs per year. In Indonesia, where the labour supply is large, duck herders take a flock of ducks to the high country during the warmer seasons and work their way down the mountainsides to the lowlands. Ducks

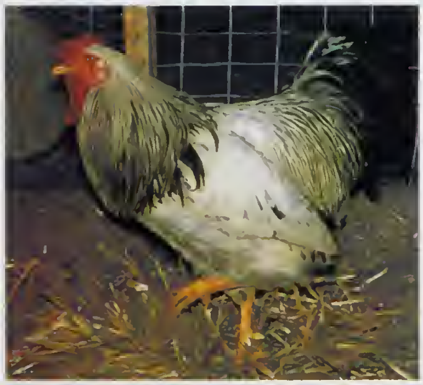
Poultry Breeds



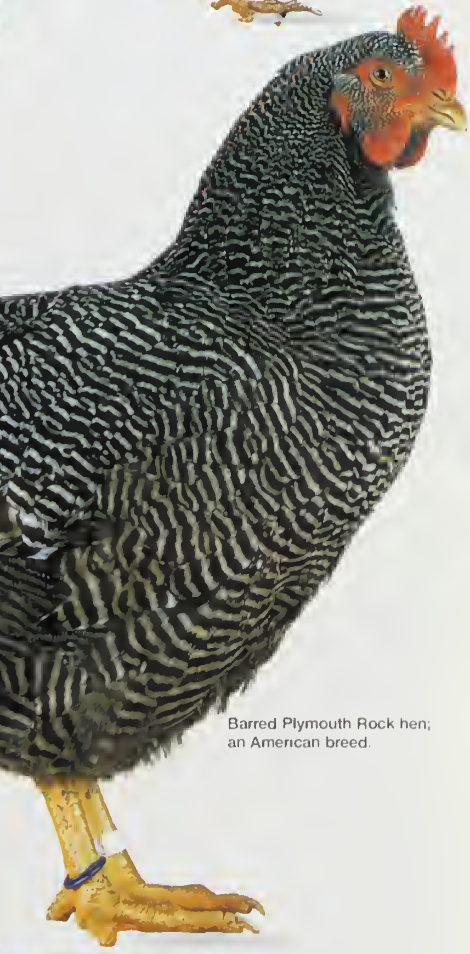
Single-comb White Leghorn cock; a Mediterranean breed.



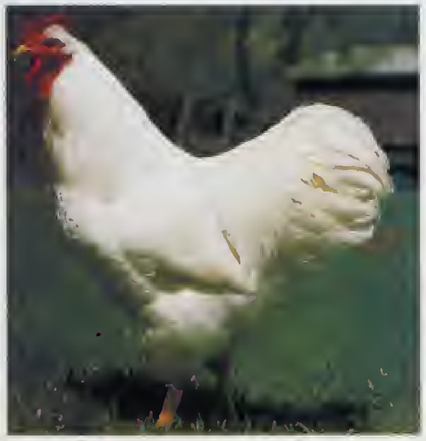
Buff Brahma cock; an Asiatic breed.



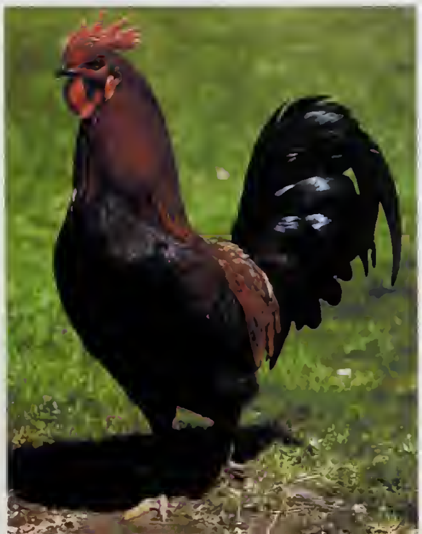
Columbian Wyandotte cock; an American breed.



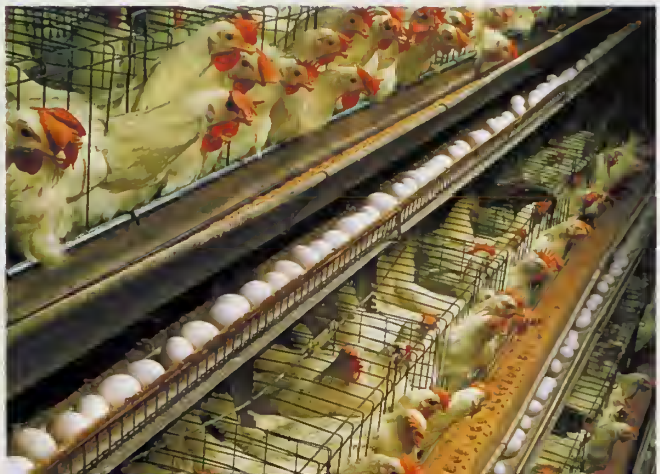
Barred Plymouth Rock hen; an American breed.



White Plymouth Rock cock; an American breed.



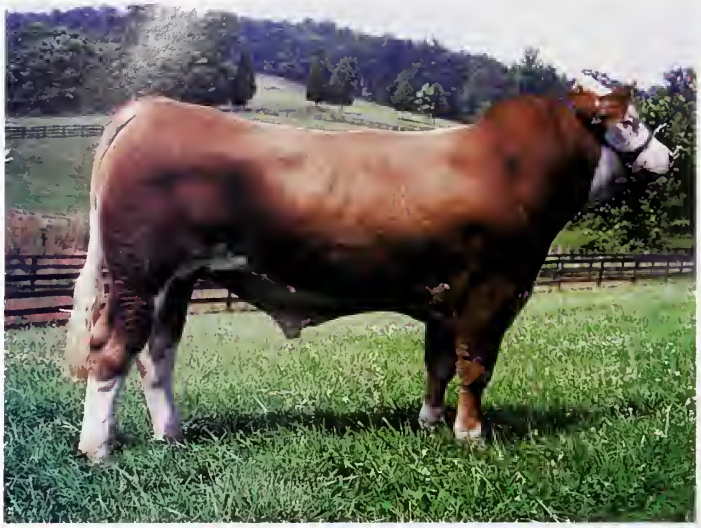
Rhode Island Red cock; an American breed.



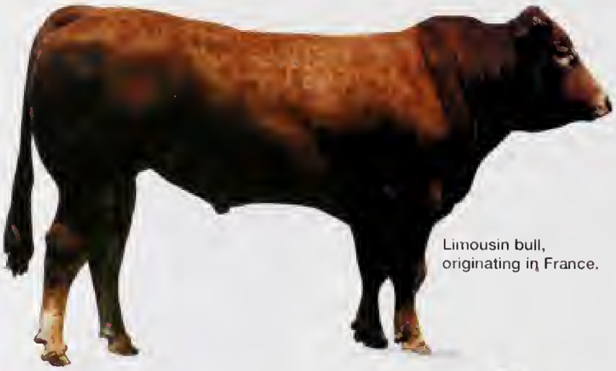
Single-comb White Leghorn hens housed for egg production in a multitiered layer house.

[illegible]
Walter [illegible] (t) [illegible]
Marshall (bottom left)
© Dr. E. R. Degginger; [illegible]
photograph by [illegible]
Lefever/Grant Heilman
Photography Inc.

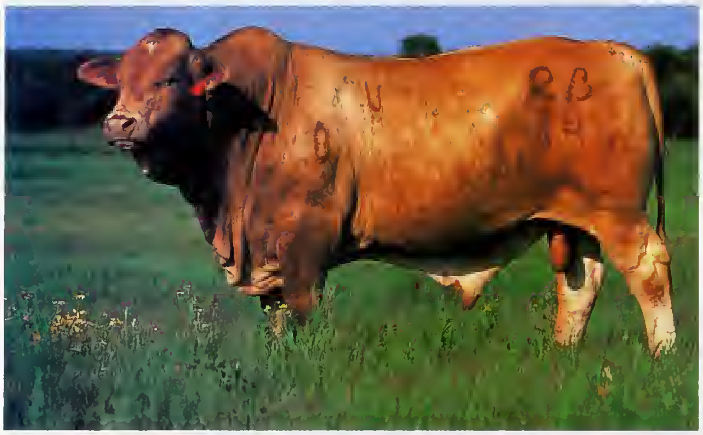
Beef Cattle Breeds



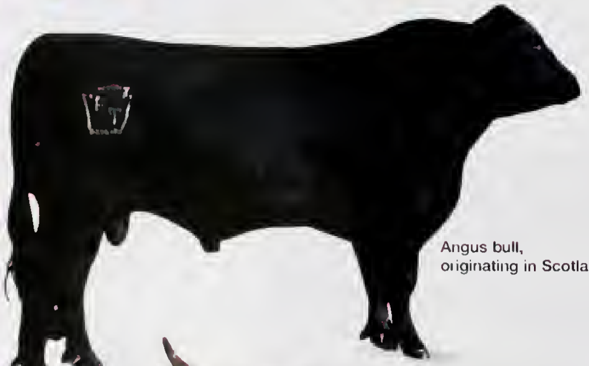
Simmental bull, originating in Switzerland.



Limousin bull, originating in France.

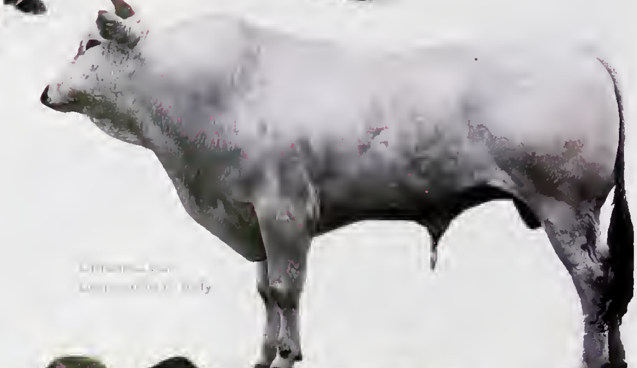


Beefmaster bull, originating in the United States.

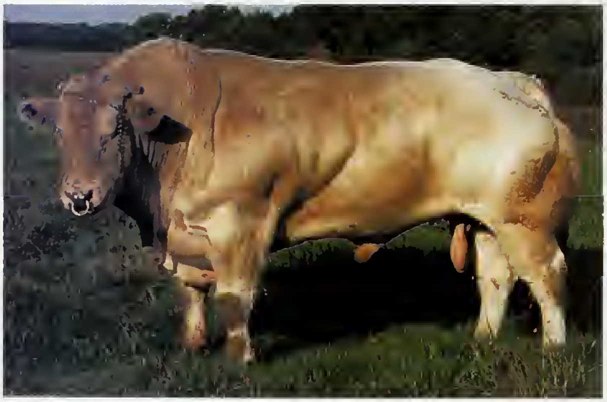


Angus bull, originating in Scotland.

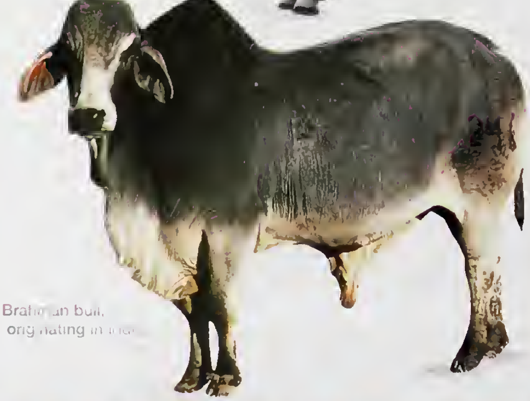
Plate 2 (Top left, top right, upper middle right) © Phil Reid Livestock Photography, (upper middle left) © B.E. Fichte, (lower middle left) Olson Family Belgian Blues, photograph, Benoit Cassart, D'Ochain, Belgium, (bottom left) © Jim Oltersdorf, (centre right) © John Colwell/Grant Heilman Photography, Inc (lower middle right) © Ronald E. Partis/Unicorn Stock Photos, (bottom right) © Grant Heilman/Grant Heilman Photography, Inc



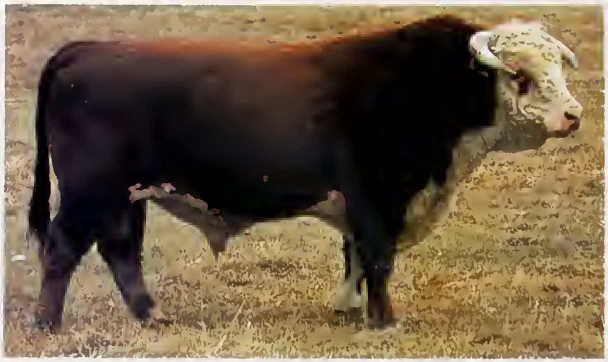
Chianina bull, originating in Italy.



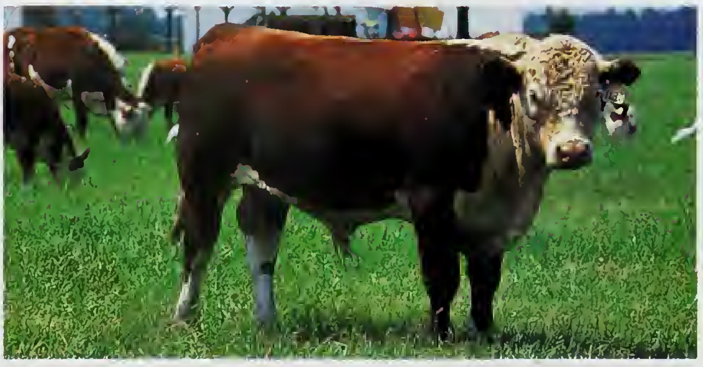
Belgian Blue bull, originating in Belgium.



Brahman bull, originating in India.

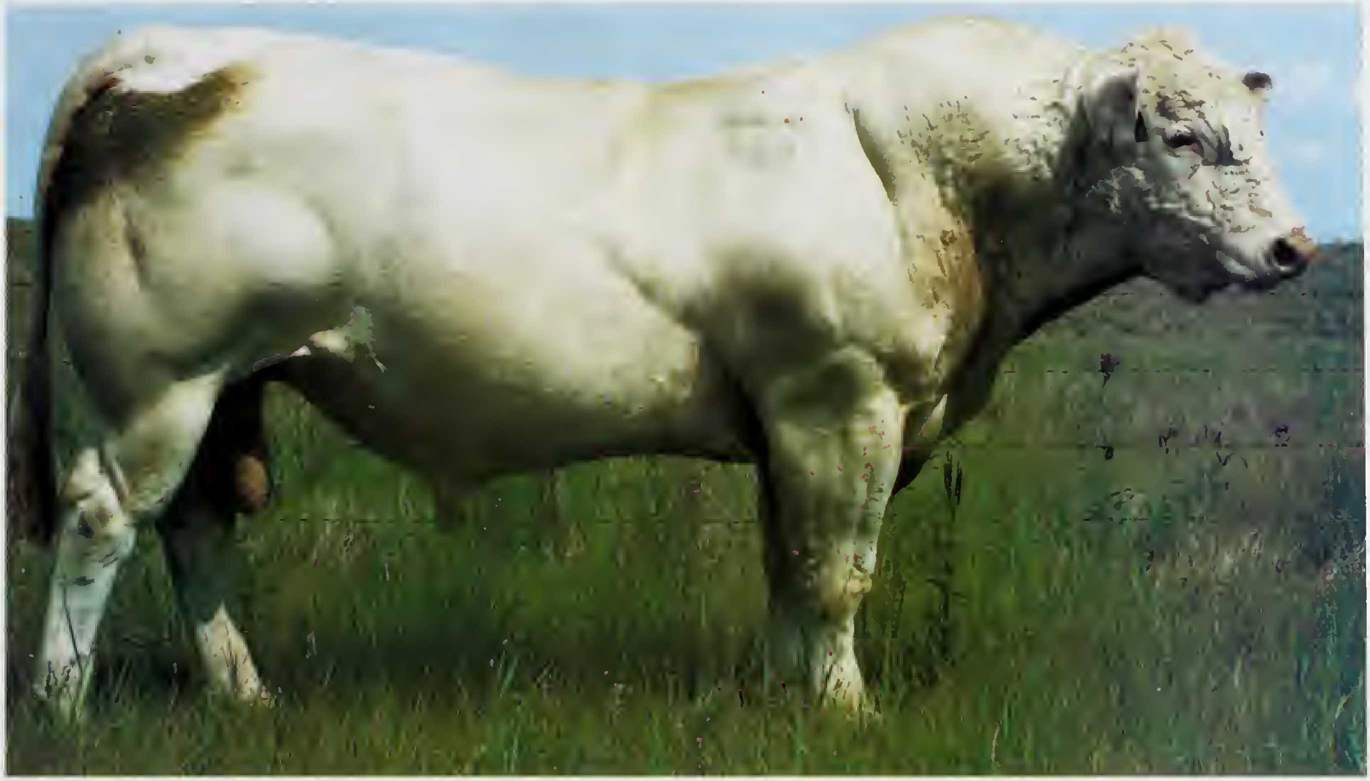


Hereford bull, originating in England.

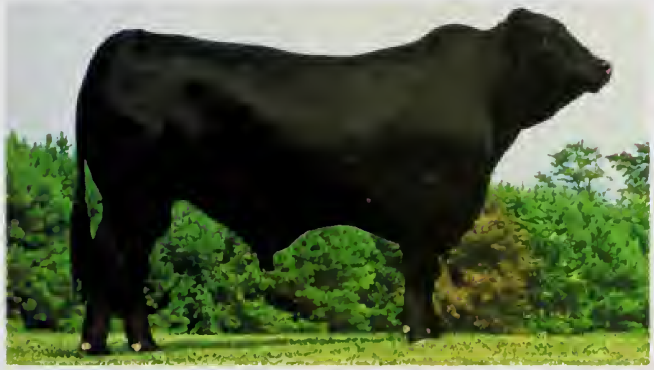


Polled Hereford bull, originating in England.

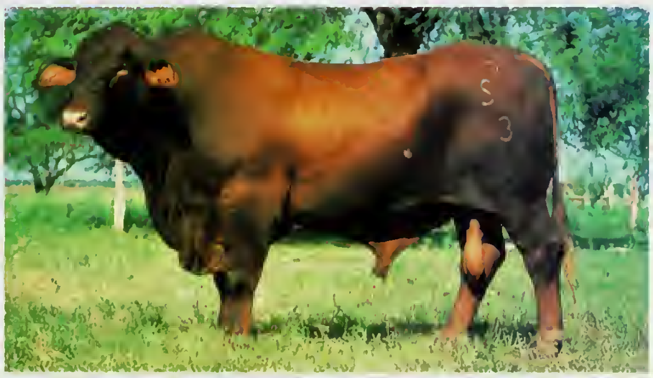
Beef Cattle Breeds



Charolais bull, originating in France

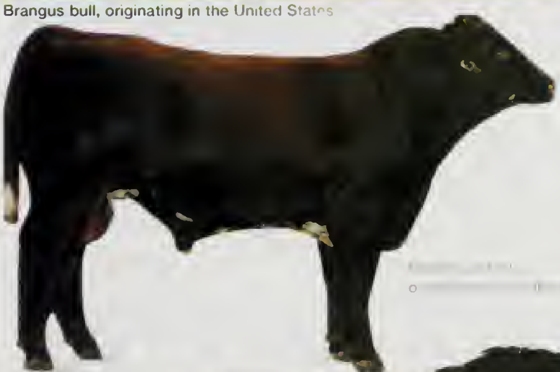


Brangus bull, originating in the United States

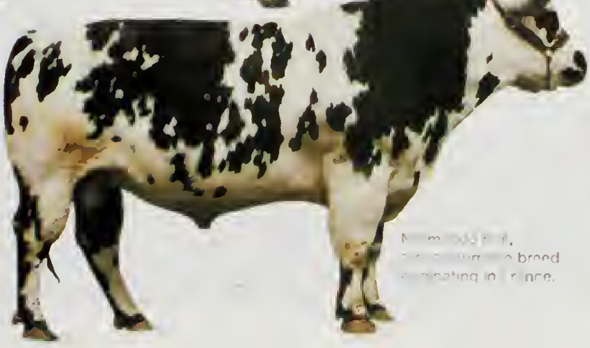


Santa Gertrudis bull, originating in the United States.

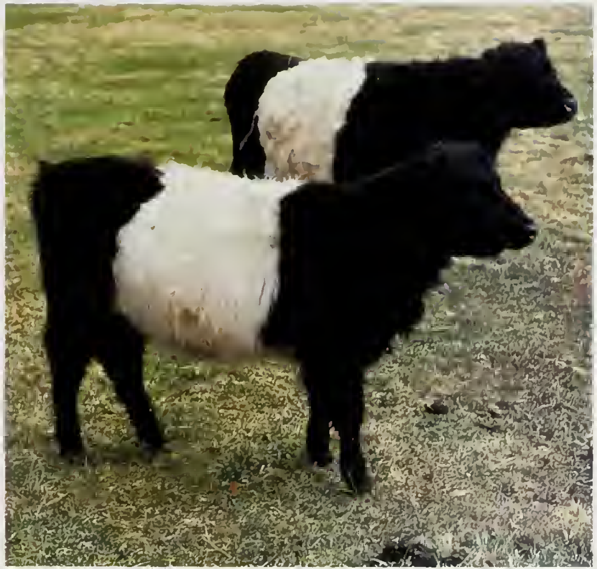
Plate 3: (Top) © American International Charolais Association, (upper middle left) © International Brangus Breeders Association, (lower middle left) © The National Livestock Exhibitor, (bottom left) © North American Normande Association, (centre right) © King Ranch Archive, (bottom right) © James Marshall



[illegible] [illegible]
o[illegible] [illegible]

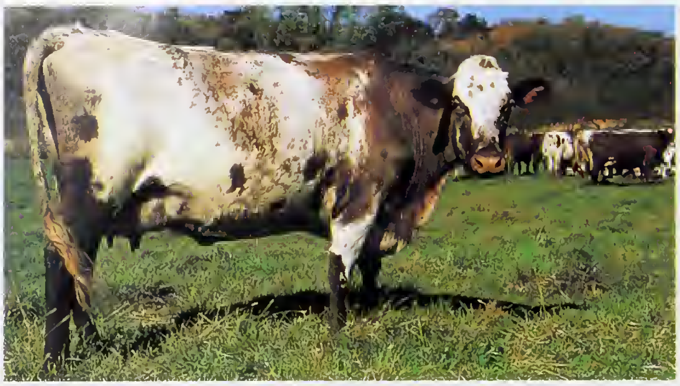


Normande bull,
a dual-purpose breed
originating in France.

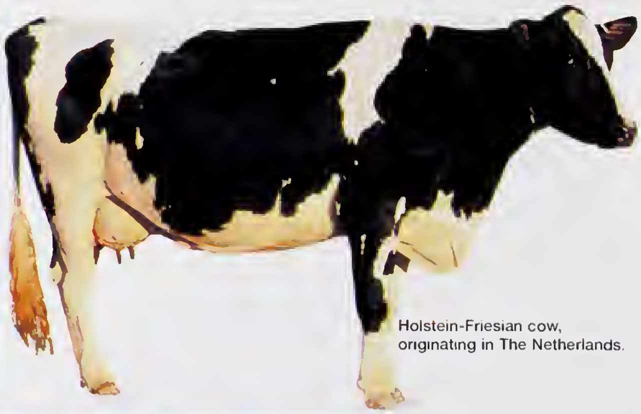


Belted Galloway, originating in Scotland.

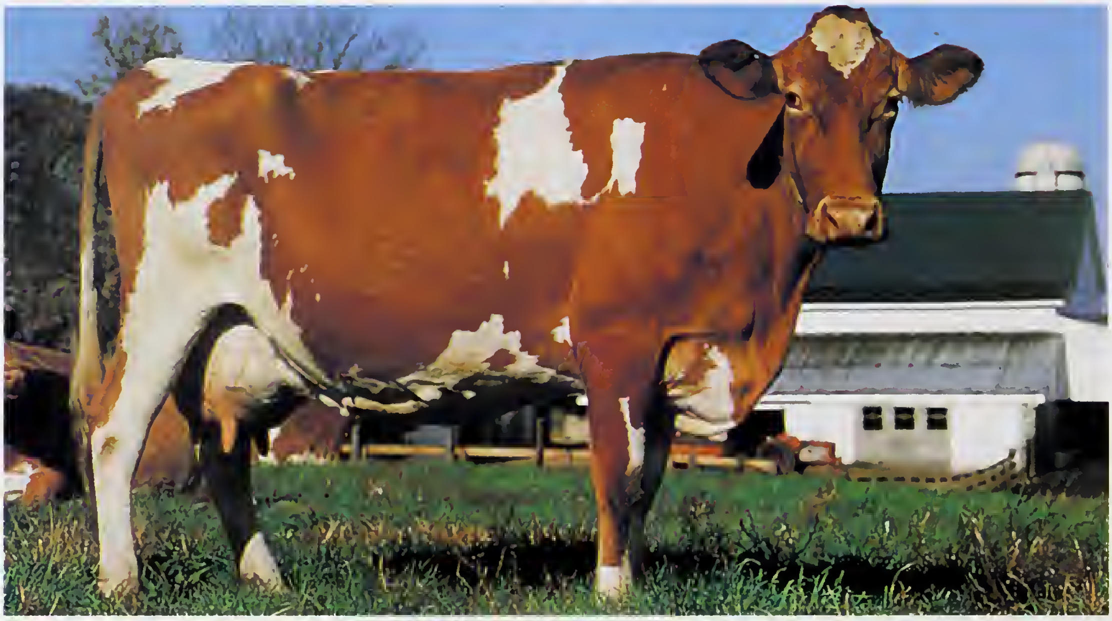
Dairy Cattle Breeds



Milking Shorthorn cow, a dual-purpose breed developed in the United States from British Shorthorn stock.

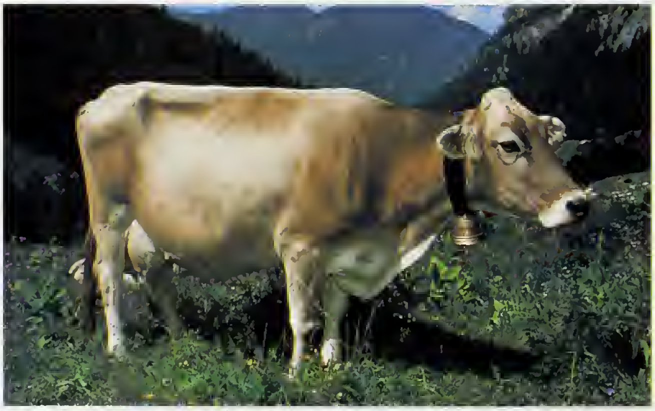


Holstein-Friesian cow, originating in The Netherlands.

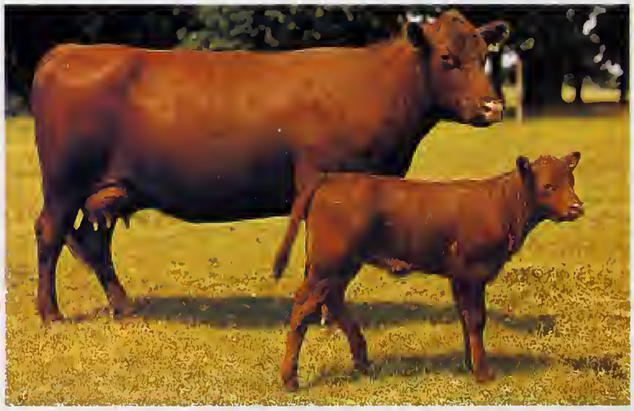


Guernsey cow, originating on the island of Guernsey.

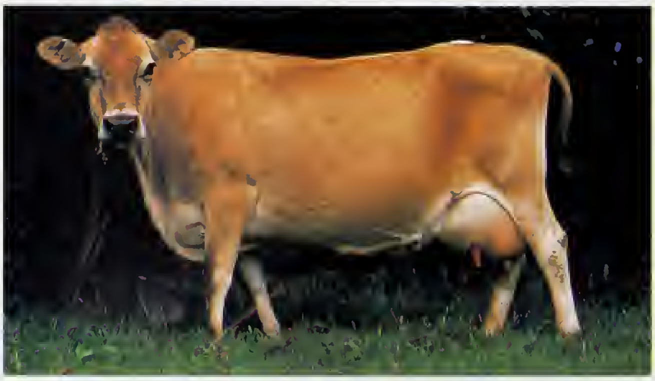
Plate 4. (Top left) © Lynn M Stone/Instock, (top right, upper middle, bottom right) © Larry Lefever/Grant Heilman Photography, Inc., (lower middle left, bottom left) © Sally Anne Thompson/Animal Photography, (lower middle right) © J C Allen and Son



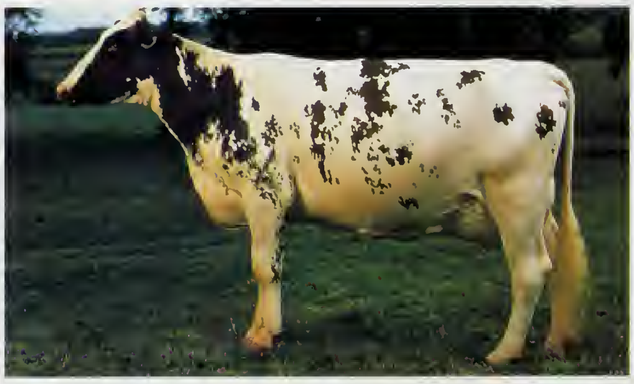
Brown Swiss cow, originating in Switzerland.



Red Poll cow and calf, a dual-purpose breed originating in England.

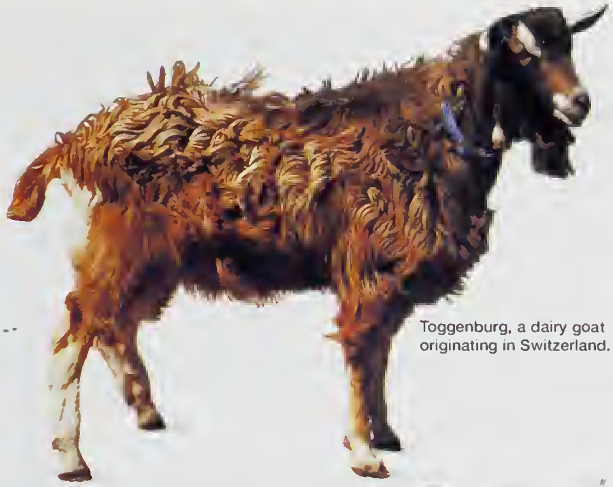


Jersey cow, originating on the island of Jersey.

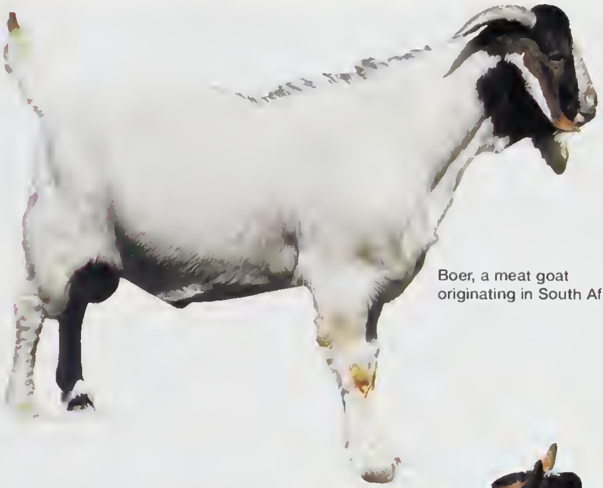


Ayrshire cow, originating in England.

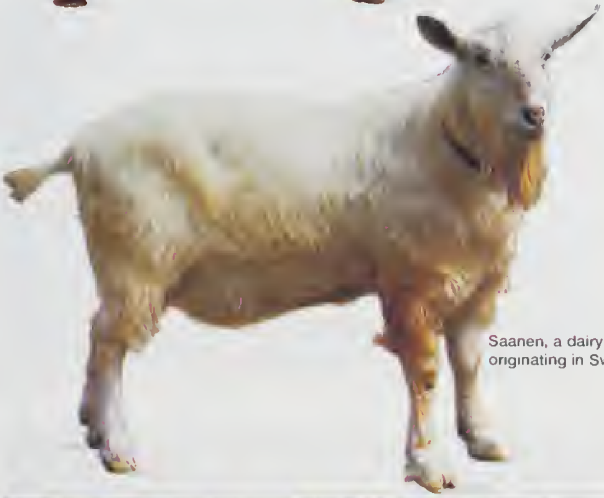
Goat Breeds



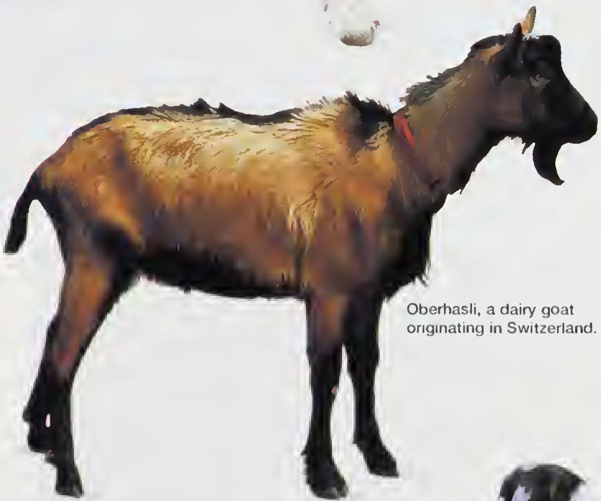
Toggenburg, a dairy goat originating in Switzerland.



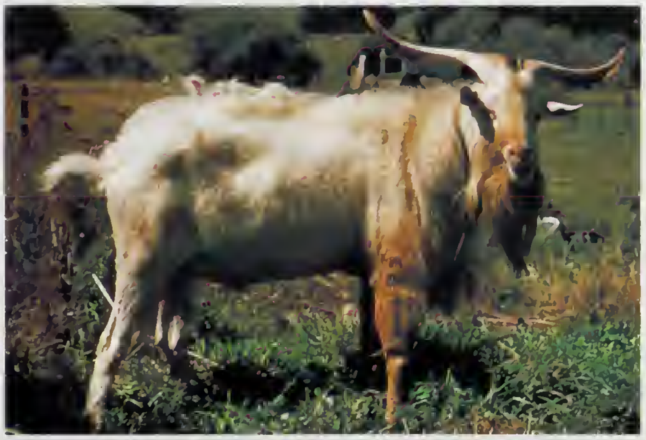
Boer, a meat goat originating in South Africa.



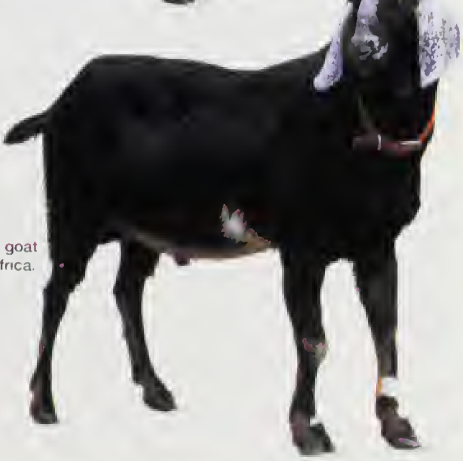
Saanen, a dairy goat originating in Switzerland.



Oberhasli, a dairy goat originating in Switzerland.

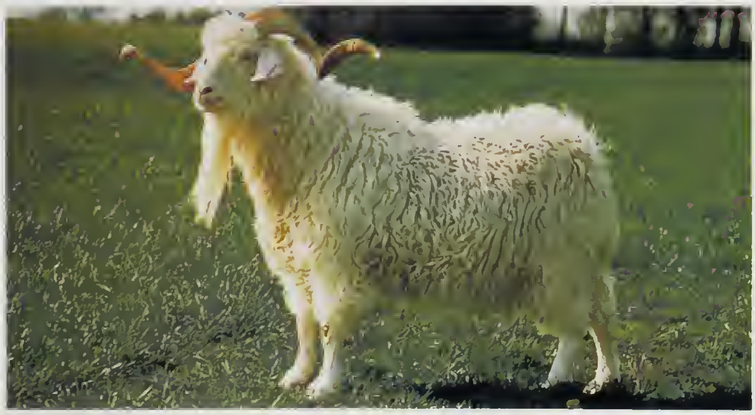


Cashmere, a wool goat originating in Asia.

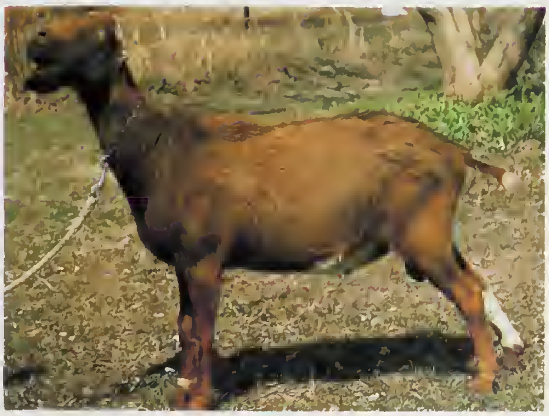


Nubian, a dairy goat originating in Africa.

Plate [illegible] (middle left) [illegible] E.R. Degginger; (bottom left) [illegible] R. Willbie/Animals [illegible] Photography; (bottom right) [illegible] John Green; all other photographs by [illegible] James Marcus



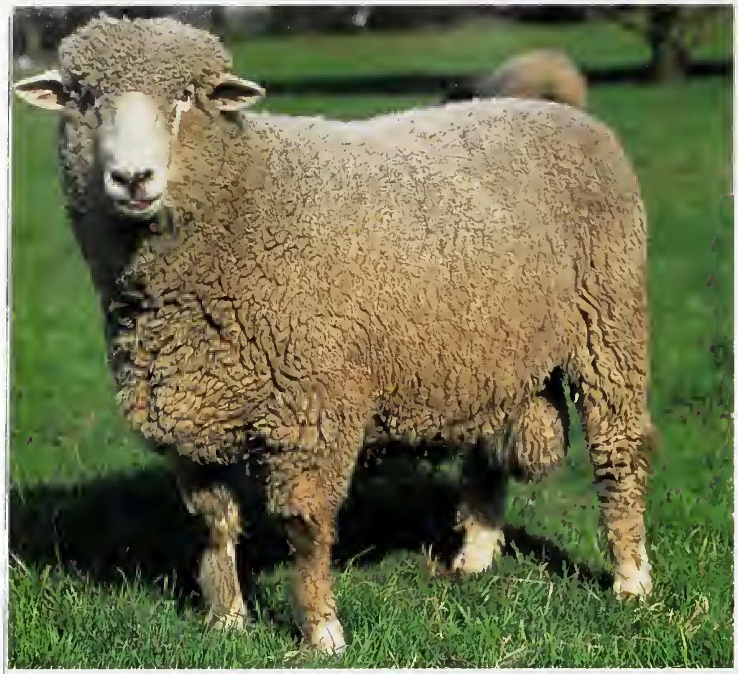
Angora, a wool goat originating in Anatolia.



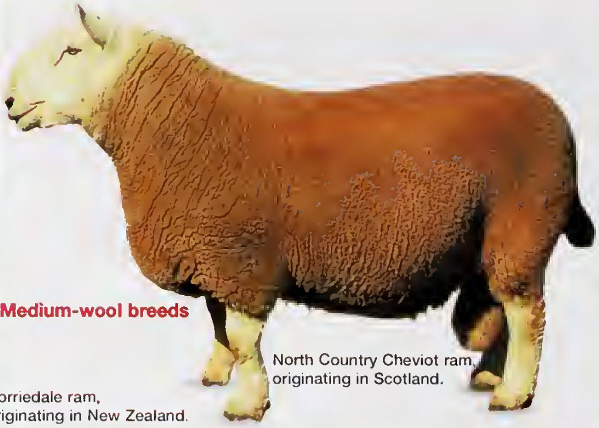
La Mancha, a dairy goat developed in the United States from Spanish stock.

Sheep Breeds

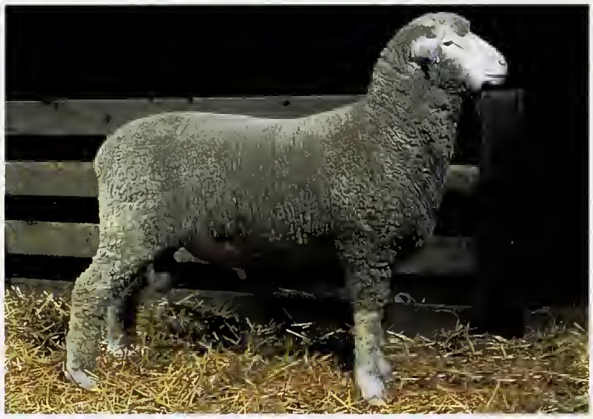
Medium-wool breeds



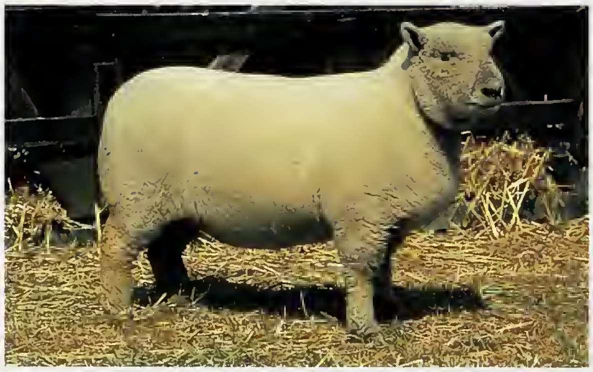
Corriedale ram, originating in New Zealand.



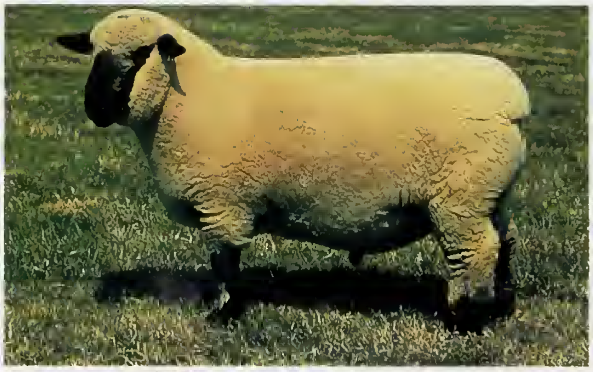
North Country Cheviot ram, originating in Scotland.



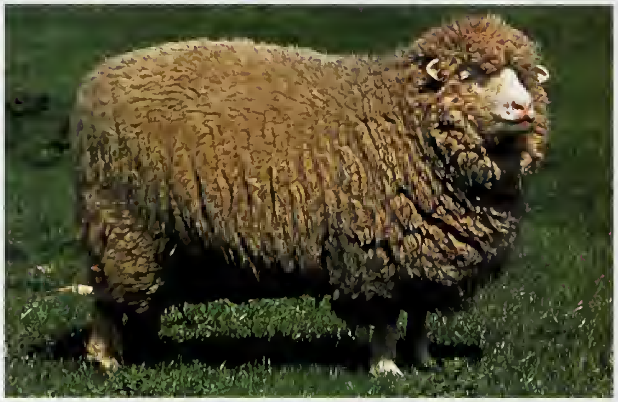
Columbia ram, originating in the United States.



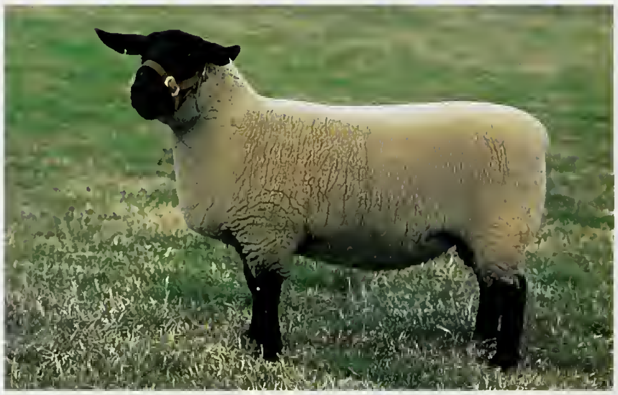
Southdown ram, originating in England.



Hampshire ram, originating in England.

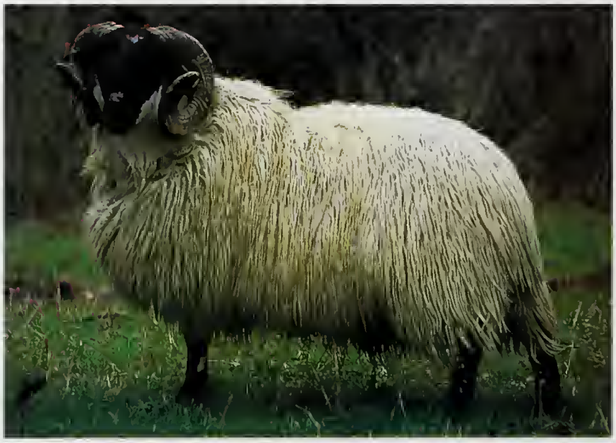


Dorset ram, originating in England.



Suffolk ram, originating in England.

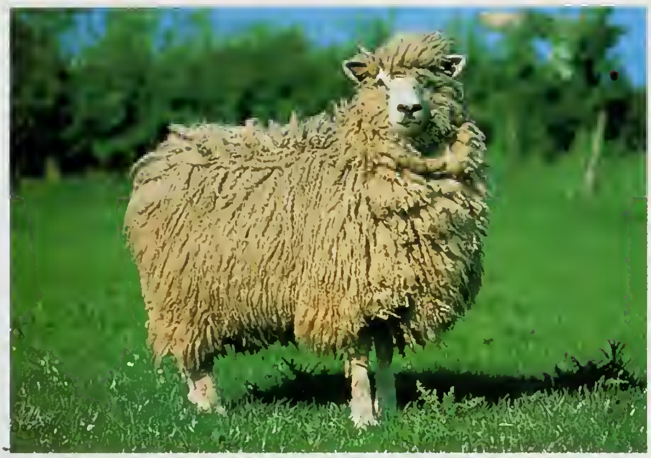
Carpet-wool breed



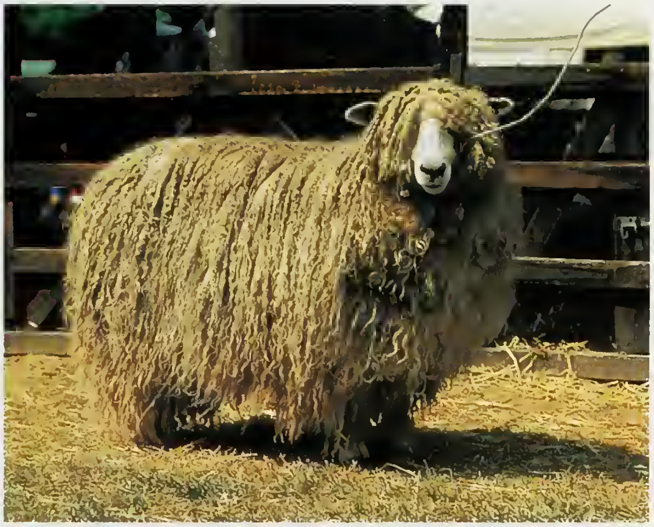
Black-Faced Highland ram, originating in Scotland.

Plate 6. (Top left) © James Marshall, (upper middle left) © John Colwell/Grant Heilman Photography, Inc., (upper middle right) © R. Willbie/Animal Photography, (bottom right) © Charmaine Walters, all other photographs by © Sally Anne Thompson/Animal Photography

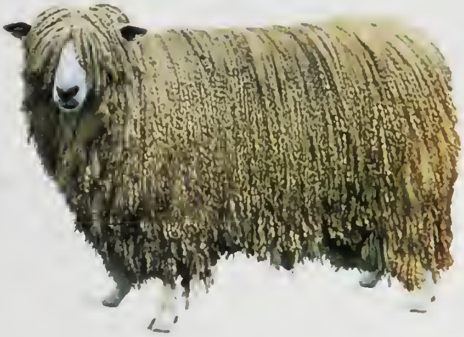
Sheep Breeds



Cotswold ewe, originating in England.

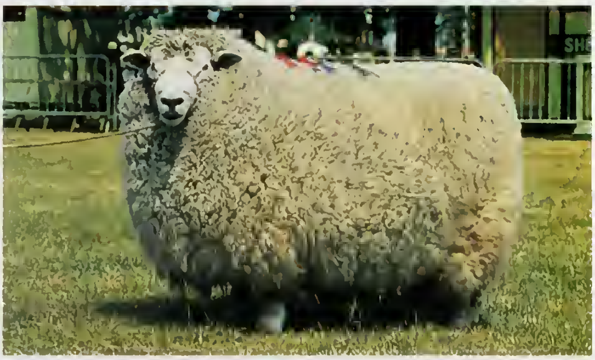


Lincoln ram, originating in England.



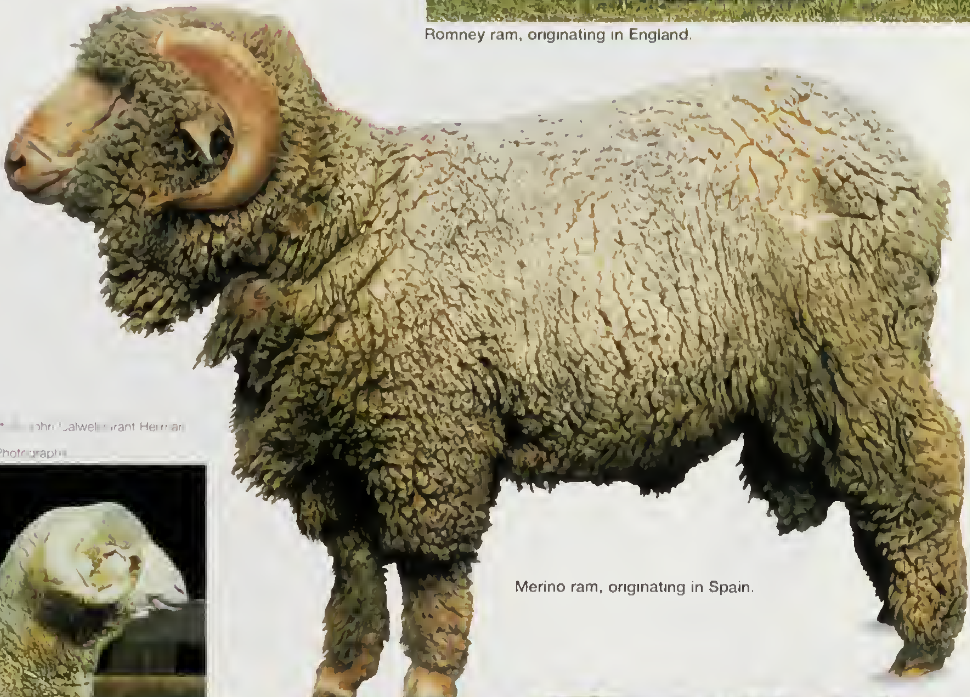
Leicester ram, originating in England.

Long-wool breeds



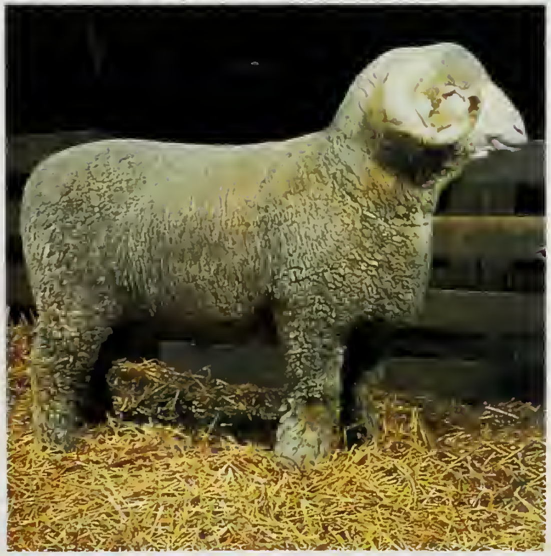
Romney ram, originating in England.

Fine-wool breeds

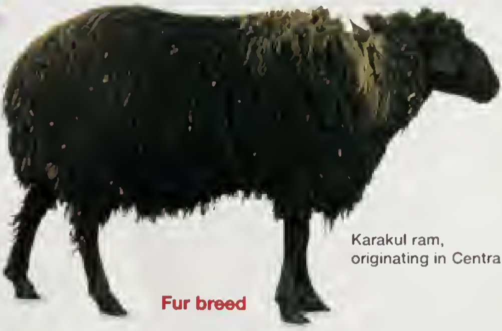


Merino ram, originating in Spain.

Plate 7: Lower middle right, © James Marshall; (bottom left) © John Colwell/Grant Heilman Photography, Inc.; (bottom right) © Charmaine Walters
all other photographs by © Sally Anne Thompson/Animal Photography



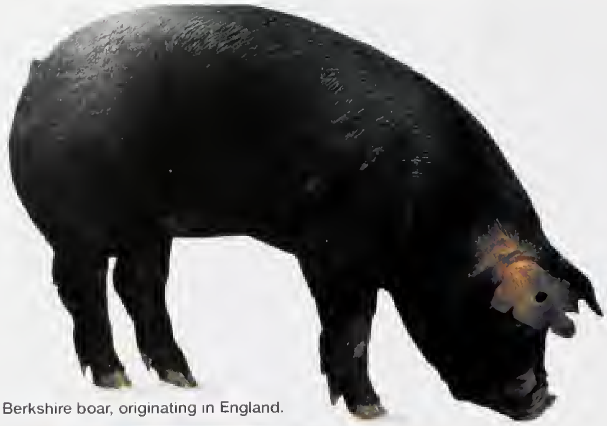
Rambouillet ram, originating in France.



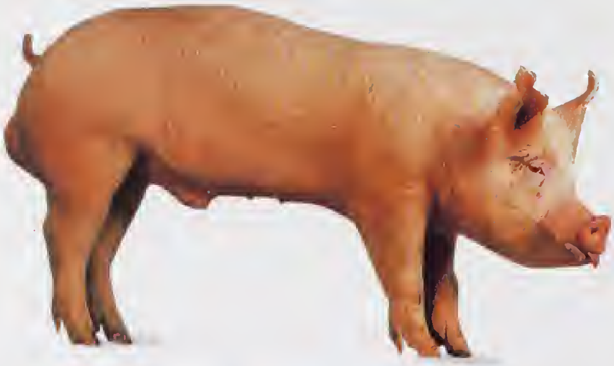
Karakul ram, originating in Central Asia.

Fur breed

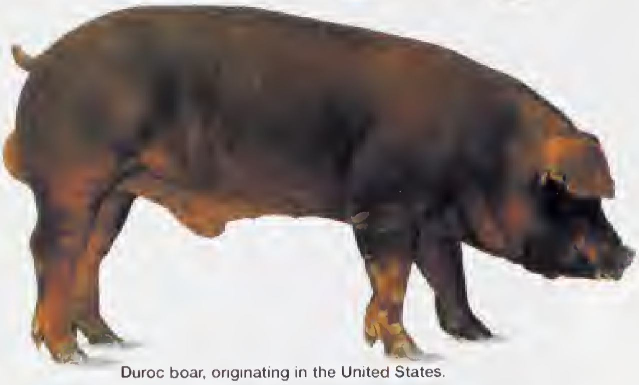
Pig Breeds



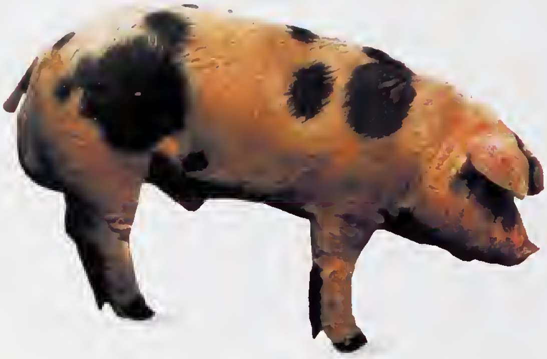
Berkshire boar, originating in England.



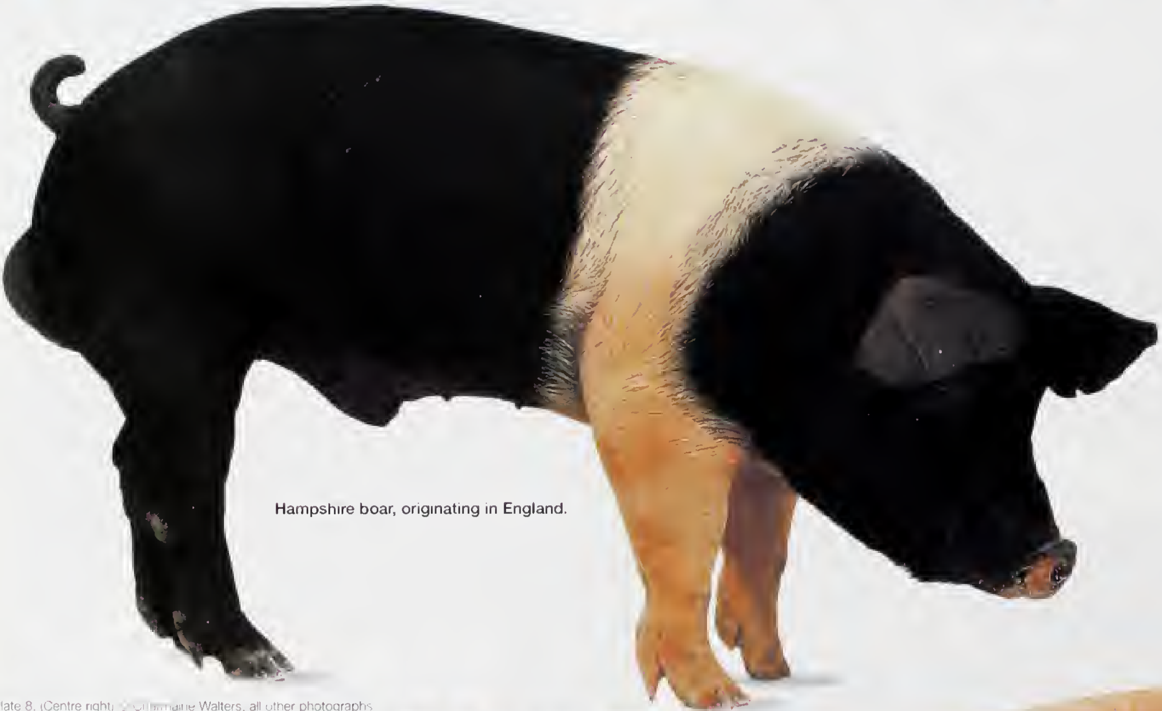
Yorkshire boar, originating in England.



Duroc boar, originating in the United States.

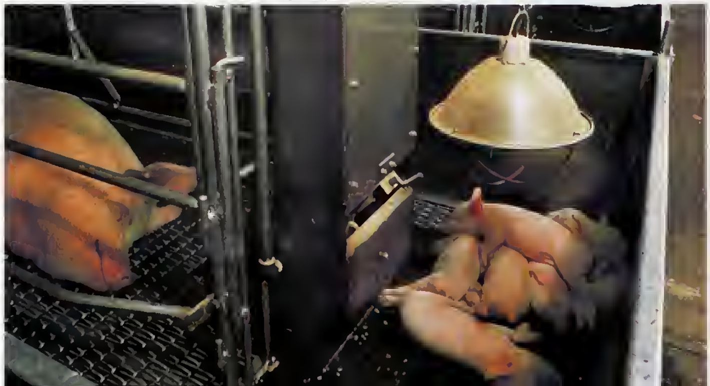


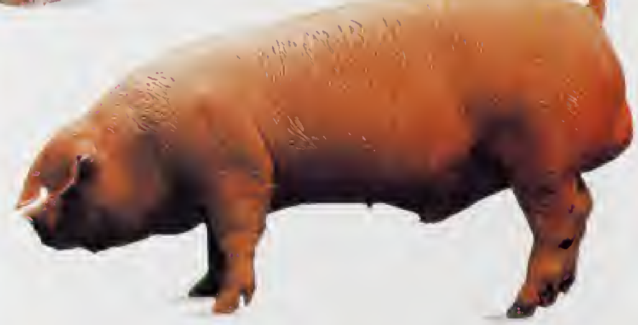
Spotted boar, originating in the United States.



Hampshire boar, originating in England.

Plate 8. (Centre right) © Charmaine Walters, all other photographs by © Larry Lefever/Grant Heilman Photography, Inc





Landrace boar, originating in Denmark.

A heat lamp warming the litter of a Yorkshire sow in a farrowing pen.

are easily transported, can be raised in close confinement, and convert some waste products and scattered grain (*e.g.*, by gleaning rice fields) to nutritious and very desirable eggs and meat. In developed countries, commercial plants have been built exclusively for duck meat production; an example is the large duckling industry of Long Island, New York. There are also local industries in The Netherlands and England, the favourite breed in England being the Aylesbury. This breed has white flesh and can reach eight pounds (3.6 kilograms) in eight weeks. The U.S. favourite is the Pekin duck, which is slightly smaller than the Aylesbury and yellow-fleshed.

Geese

Goose raising is a minor farm enterprise in practically all countries, but in Germany, Austria, some eastern European countries (notably Poland), parts of France, and locally elsewhere, there is important commercial goose production. The two outstanding meat breeds are the Toulouse, predominantly gray in colour, and the Embden (or Emden), which is white. Geese do not appear to have attracted the attention of geneticists on the same scale as the meat chicken and the turkey, and no change in the goose industry comparable to that in the others has occurred or seems to be in prospect. In some commercial plants, geese are fattened by a special process resulting in a considerable enlargement of their livers, which are sold as a delicacy, pâté de foie gras.

Guinea fowl and squabs. Guinea fowl are raised as a sideline on a few farms in many countries, and eaten as gourmet items. In Italy there is a fairly extensive industry. There the birds are raised in yards with open-fronted shelters. In England, guinea fowl are marketed at 16–18 weeks of age and in the U.S. at about 10–12 weeks. The market weight is usually about 2½–3½ pounds, but food conversion is poor.

Pigeons are raised not only as messengers and for sport but also for the meat of their squabs (nestlings), also a gourmet item. Squab production, carried on locally, is rare in most countries with established poultry industries.

POULTRY DISEASES

Poultry are quite susceptible to a number of diseases; some of the more common are fowl typhoid, pullorum, fowl cholera, chronic respiratory disease, infectious sinusitis, infectious coryza, avian infectious hepatitis, infectious synovitis, bluecomb, Newcastle disease, fowl pox, avian leukosis complex, coccidiosis, blackhead, infectious laryngotracheitis, infectious bronchitis, and erysipelas. Strict sanitary precautions, the intelligent use of antibiotics and vaccines, and the widespread use of cages for layers and confinement rearing for broilers have made it possible to effect satisfactory disease control.

Parasitic diseases of poultry, including hexamitiasis of turkeys, are caused by roundworms, tapeworms, lice, and mites. Again, modern methods of sanitation, prevention, and treatment provide excellent control. (W.P.G.)

Beekeeping

Beekeepıng is the care and management of colonies of honeybees. They are kept for their honey and other products or their services as pollinators of fruit and vegetable blossoms or as a hobby. The practice is widespread: honeybees are kept in large cities and villages, on farms and rangelands, in forests and deserts, from the Arctic and Antarctic to the Equator.

In antiquity people knew that bees produce delicious honey, that they sting, and that they increase their numbers by swarming. By the 17th century they had learned the value of smoke in controlling them and had developed the screen veil as protection against stings. From the 17th to the 19th century, the key discoveries upon which modern beekeeping is founded were made. These included the mystery of the queen bee as the mother of nearly all the occupants of the hive, her curious mating technique, parthenogenetic development, the movable frame hives, and the fact that bees rear a new queen if the old one disappears.

Given this knowledge people were able to divide a colony instead of relying on natural swarming. Then the development of the wax-comb foundation, the starter comb on which bees build straight, easily-handled combs, and the discovery that honey can be centrifuged or extracted from them and the combs reused, paved the way for large-scale honey production and modern commercial beekeeping. The identification of bee diseases and their control with drugs, the value of pollen and pollen substitutes in producing strong colonies, and the artificial insemination of queens have increased the honey-production efficiency of colonies.

HONEYBEES AND THEIR COLONIES

Honeybees. Honeybees belong to the order Hymenoptera and to one of the *Apis* species. (For a complete discussion of honeybees, see the article INSECTS: *Hymenoptera.*) Honeybees are social insects noted for providing their nests with large amounts of honey. A colony of honeybees is a highly complex cluster of individuals that functions virtually as a single organism. It usually consists of the queen bee, a fertilized female capable of laying a thousand or more eggs per day; from a few to 60,000 sexually undeveloped females, the worker bees; and from none to 1,000 male bees, or drones. The female of most species of bees is equipped with a venomous sting.

The queen

Honeybees collect nectar, a sugary solution, from nectaries in blossoms and sometimes from nectaries on the leaves or stems of plants. Nectar may consist of 50 to 80 percent water, but when the bees convert it into honey it will contain only about 16 to 18 percent water. Sometimes they collect honeydew, an exudate from certain plant-sucking insects, and store it as honey. The primary carbohydrate diet of bees is honey. They also collect pollen, the dustlike male element, from the anthers of flowers. Pollen provides the essential proteins necessary for the rearing of young bees. In the act of collecting nectar and pollen to provision the nest, the bees pollinate the flowers they visit. Honeybees also collect propolis, a resinous material from buds of trees, for sealing cracks in the hive or for covering foreign objects in the hive that they cannot remove. They collect water to air-condition the hive and to dilute the honey when they consume it.

A populous colony in a desirable location may, in a year's time, collect and carry into the hive as much as 1,000 pounds (450 kilograms) of nectar, water, and pollen.

Honeycomb

Bees secrete beeswax in tiny flakes on the underside of the abdomen and mold it into honeycomb, thin-walled, back-to-back, six-sided cells. The use of the cell varies depending on the needs of the colony. Honey or pollen may be stored in some cells, while the queen lays eggs, normally one per cell, in others. The area where the bees develop from the eggs is called the broodnest. Generally, honey is stored toward the top of the combs and pollen in cells around the broodnest below the honey.

The bees maintain a uniform temperature of about 93° F (34° C) in the broodnest regardless of outside temperature. The colony can survive daily maximum temperatures of 120° F (49° C) if water is available with which they can air-condition the cluster. When the temperature falls below about 57° F (14° C) the bees cease flying, form a tight cluster to conserve heat, and await the return of warm weather. They can survive for several weeks in temperatures of −50° F (−46° C).

When summer flowers bloom in profusion, the queen's egg-laying is stimulated, the cluster expands, and honey accumulates in the combs. When the large number of young bees emerge, the domicile becomes crowded.

Swarming. When the colony becomes crowded with adult bees and there are insufficient cells in which the queen can lay large numbers of eggs, the worker bees select a dozen or so tiny larvae that would otherwise develop into worker bees. These larvae are fed copiously with royal jelly, a whitish food with the consistency of mayonnaise, produced by certain brood-food glands in the heads of the worker bees. The cell in which the larva is developing is drawn out downward and enlarged to permit development of the queen. Shortly before these virgin queens emerge as adults from their queen cells, the mother queen departs from the beehive with the swarm. Swarming usually occurs during the middle of a warm day, when the queen and a portion of the worker bees (usually from 5,000 to 25,000) suddenly swirl out of the hive and into the air. After a few minutes' flight, the queen alights, preferably on a branch of a tree but sometimes on a roof, a parked automobile, or even a fire hydrant. All the bees settle into a tight cluster around her while a handful of scouts reconnoitre a new homesite.

When the scout bees have located a new domicile, the cluster breaks, the swarm takes to the air and in a swirling mass proceeds to the new home. Swarming is the bees' natural method of propagation or increase.

Queen bee. Back in the parent colony, the first queen to emerge after the mother queen departs with the swarm immediately attempts to destroy the others. If two or more emerge at the same time, they fight to the death. When the surviving virgin is about a week old, she soars off on her mating flight; she frequently mates with more than one drone while in the air. She may repeat the mating flights for two or three successive days, after which she begins egg laying. She rarely ever leaves the hive again except with a swarm. Normally, sufficient sperm are stored in her sperm pouch, or spermatheca, to fertilize all the eggs she will lay for the rest of her life. The drones die in the act of mating.

The queen can live up to five years, although many beekeepers replace the queen every year or two. If she is accidentally killed or begins to falter in her egg-laying efficiency, the worker bees will rear a "supersedure" queen that will mate and begin egg laying without a swarm emerging. She ignores the mother queen, who soon disappears from the colony.

Worker bees. Worker bees live about six weeks during the active season but may live for several months if they emerge as adults in the fall and spend the winter in the cluster. As the name implies, worker bees do all of the work of the hive, except the egg laying.

Drones. Drones are reared only when the colony is populous and there are plentiful sources of nectar and pollen. They usually live a few weeks, but are driven from the hive to perish when fall or an extended period of adversity comes upon the colony. The only duty of the drone is to mate with the queen.

The queen can lay drone (unfertilized) eggs in the drone cells. If she is not allowed to mate or if her supply of sperm is exhausted, she will lay unfertilized eggs in worker cells. The development of unfertilized eggs into adult drones is known as parthenogenesis. Occasionally a colony may become queenless and unable to develop another queen. Then, some of the worker bees begin to lay eggs, often several to a cell, and these develop into drones. A colony that has developed laying workers is difficult to requeen with a laying queen.

COLONY MANIPULATION

The yearly work cycle. The beekeeper's year starts in early fall. At that time he requeens the colonies whose queens are not producing adequate amounts of brood and makes sure that each colony has sufficient stores: at least 50 pounds (22 kilograms) of honey and several frames filled with pollen. Some beekeepers also feed the drug fumagillin to reduce possible damage to the adult bees by nosema disease (see below *Disease and pest control*). The colonies need a sunny exposure and protection from cold winds. Some beekeepers in northern and mountainous areas wrap their colonies with insulating material in winter. A few beekeepers kill their bees in the fall, harvest the honey, store the empty equipment, then restock with a two- or three-pound (0.8- or 1.4-kilogram) package of bees and a young queen the following spring.

Wintering

If the colonies are well prepared in the fall they need little attention during the winter. But in early spring, an examination of the colonies by the beekeeper is important. Frequently, strong colonies exhaust their food supply and starve only a few days before flowers begin to bloom in abundance. Only a few pounds of sugar syrup, 50-50 sugar water, or a honey-filled comb from another more prosperous colony might save such a starving colony. Again fumagillin may be fed to the colony, and some beekeepers also feed a cake of pollen substitute or pollen supplement. Honey is not fed to the colonies unless the beekeeper is sure about its source. Honey from colonies affected by the brood disease American foulbrood could infect his colonies and cause a serious loss.

As the spring season advances, the cluster size increases from the low population of 10,000 to 20,000 bees that survived the winter. To accommodate the increased size of the cluster and broodnest, the keeper adds more supers, or boxes of combs. If the combs are so manipulated that the queen can continually expand her egglaying area upward, the colony is unlikely to swarm. This can be achieved by placing empty combs or combs in which brood is about ready to emerge at the top of the cluster and combs filled with eggs or young brood toward the lower part of the broodnest. The beekeeper wants the colony to reach its peak of population, 50,000 to 60,000 bees, at the beginning of the major nectar flow.

The bees in a swarm, having departed the hive with a full stomach of honey, rarely sting. The usual way to capture them is to place a hive or upturned box beneath or nearby, then shake or smoke the bees to force the queen and a majority of the bees into it. The others follow. After the swarm is safely inside the box it can be removed to a permanent location.

Regulations governing the keeping of bees usually require the bees to be kept in hives with movable combs. If the bees are captured in a box they are generally transferred into a movable-frame hive within a few days so the new honey and comb will not be lost in the transfer.

Requeening a colony. When a beekeeper requeens a colony, he removes the failing or otherwise undesirable queen and places a new one in a screen cage in the broodnest. After a few days the colony becomes adjusted to her and she can be released from the cage. A strange queen placed in the cluster without this temporary protection usually will be killed at once by the workers. Queens usually are shipped in individual cages of about three cubic inches (50 cubic centimetres) with about half a dozen attendant bees and a ball of specially prepared sugar candy plugging one end of the cage. When the cage is placed in the hive, the bees from both sides eat the candy. By the time the candy is consumed and the bees reach each other, their odours have become indistinguishable, the queen emerges from the cage into the colony and begins her egglaying duties.

Beekeeping equipment. Standard tools of the beekeeper are: the smoker to quell the bees; a veil to protect the face; gloves for the novice or the person sensitive to stings; a blunt steel blade called a hive tool, for separating the frames and other hive parts for examination; the uncapping knife, for opening the cells of honey; and the extractor, for centrifuging the honey from the cells.

By courtesy of the U.S. Department of Agriculture

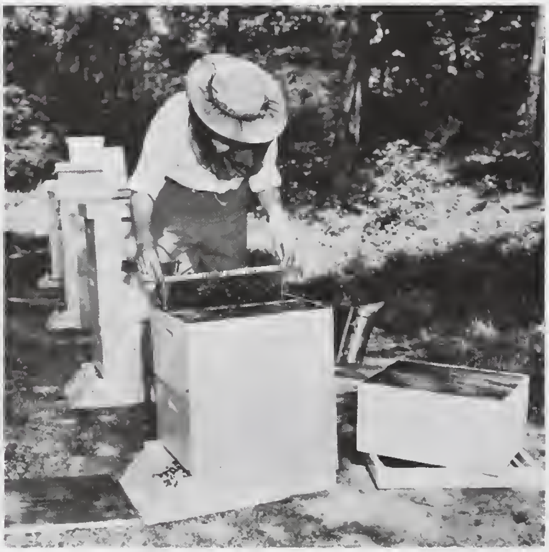


Figure 15: A beekeeper, wearing a veil and holding a hive tool, lifts a frame from a super. A smoker is attached to the side of the hive and a super has been removed.

Bee stings. The worker bee sting is barbed, and in the act of stinging it is torn from the bee. It has a venom-filled poison sac and muscles attached that continue to work the sting deeper into the flesh for several minutes and increase the amount of venom injected. To prevent this, the sting should be scraped loose (rather than grasped and pulled out) at once. Bee stings are painful, and no one becomes immune to the pain. Immunity to the swelling is usually built up after a few stings, however.

Reactions to stings

Normal reaction to a bee sting is immediate, intense pain at the site of the sting. This lasts for a minute or two and is followed by a reddening, which may spread an inch or more. Swelling may not become apparent until the following day. Occasionally, acute allergic reactions develop from a sting, usually with persons who have other allergic problems. Such a reaction becomes evident in less than an hour and may consist of extreme difficulty in breathing, heart irregularity, shock, splotched skin, and speech difficulty. Such persons should obtain the services of a medical doctor immediately.

BEE PRODUCTS

Honey production. Honey is marketed in several different forms: liquid honey, comb honey, and creamed honey. Sometimes the predominant floral type from which the honey was collected is indicated.

Liquid honey. If liquid (strained, extracted) honey is desired, additional supers are added directly above the brood nest. When one is largely filled, it is raised and another is placed underneath. This may continue until several have been filled, each holding from 30 to 50 pounds (14 to 23 kilograms), or until the nectar flow has ended. After the bees have evaporated the water until the honey is of the desired consistency and sealed in the cells, the combs are removed, the cells uncapped with the uncapping knife, and the honey extracted. The removed honey is immediately heated to about 140° F (60° C), which thins it and destroys yeasts that can cause fermentation. It is then strained of wax particles and pollen grains, cooled rapidly, and packaged for market.

Comb honey. In production of honey in the comb, or comb honey, extreme care is necessary to prevent the bees' swarming. The colony must be strong, and the bees must be crowded into the smallest space they will tolerate without swarming. New frames or sections of a frame with extra-thin foundation wax, added at exactly the right time for the bees to fill without destroying them, are placed directly above the brood nest. The bees must fill and seal the new comb to permit removal within a few days, or it will be of inferior quality. As rapidly as sections are removed, new sections are added, until the nectar flow subsides; then these are removed and the colony given combs to store its honey for the winter.

Creamed honey. Almost all honey will granulate or turn to sugar. Such honey can be liquefied without materially affecting its quality by placing the container in water heated to about 150° F (66° C). Liquid and granulated honey is sometimes blended, homogenized, and held at a cool temperature, which speeds uniformly fine granulation. If properly processed, the granules will be extremely fine; the honey, which has a smooth, creamy appearance, is referred to as creamed honey.

Floral types. Some honeys are sold by floral type; that is, they are given the name of the predominant flowers visited by the bees when they accumulated the honey. The beekeeper has no way to direct the bees to a particular source of food but through experience learns which plants are the major sources of honey. Different flowers produce different colours and flavours of honey. It may be heavy-bodied or thin-bodied, dark or light, mild-flavoured or strong-flavoured. Most honey has been blended by the beekeeper to a standard grade that can be supplied and marketed year after year.

Colour and flavour varieties

World honey production statistics. World production of honey was about 1,950,000,000 pounds (884,000,000 kilograms) in 1981. North America produced about 390,000,000 pounds (180,000,000 kilograms); the U.S.S.R., about 420,000,000 pounds (190,000,000 kilograms); and the remaining countries of the world, about 1,140,000,000 pounds (517,000,000 kilograms). The largest exporting countries were mainland China, Mexico, Argentina, and Australia. Although honey production per colony in the United States amounted to about 44 pounds (20 kilograms), when the colonies are properly manipulated and in good locations they frequently produce several times this amount. An average annual production of several hundred pounds per colony has been reported for a small isolated area of southwestern Australia.

Beeswax. Beeswax is a by-product of beekeeping in most areas. When beekeepers uncap or break honeycombs or have unusable combs, they try to salvage the beeswax. First, they recover as much honey from the combs as possible by drainage or extraction. Then they place the material in water heated to slightly over 145° F (63° C). This melts the wax, which rises to the surface. After it cools and hardens, the cake of wax is removed and refined for reuse in comb foundation. Beeswax has many other uses: in quality candles, cosmetics, agriculture, art, and industry. In some areas bees are manipulated primarily for wax production. Wax is a highly stable commodity that can be transported long distances under unfavourable conditions without damage.

Bees reared for sale. Queens are reared for sale to other beekeepers for requeening established colonies or for adding to a two- or three-pound (0.9- or 1.4-kilogram) package of 8,000 to 10,000 live bees to form new colonies or replenish weak ones. The queens are produced when the beekeeper cages the reigning queen in a colony, then inserts into the cluster from 30 to 60 queen cell bases into which young (one-day-old) worker larvae have been transferred. More than 1,000,000 queens are produced in this way and sold each year in the United States. Queens can be artificially inseminated with sperm from drones of a known source, but most beekeepers let the queens mate naturally.

The live bees are shaken from the combs of the colony through a funnel into screen-wire cages. About 500 tons of live bees are produced for sale annually in the United States, primarily in the southeastern states and California. Several tons are shipped annually from the United States to foreign countries, primarily to Canada.

Pollination. The greatest value of bees is in their service as pollinators. Some 90 crops grown in the United States alone are dependent on insect pollination, performed primarily by the honeybee. The average colony of bees is worth from 20 to 40 times as much in the pollination of crops as it is in the production of honey. The value of bees in the pollination of ornamental plants has never been calculated. Bees are also valuable in the pollination of some forest and range plants that produce seeds on which birds and other wildlife feed.

Bees in pollination

When bees are used in the pollination of crops, the beekeeper places the colonies within or adjacent to the field to be pollinated. The majority of the roughly 1,000,000 colonies that are used for pollination are used in alfalfa-seed fields and almond and apple orchards. The colonies are distributed at the rate of two or more per acre in groups every 0.1 mile (0.16 kilometre) throughout alfalfa fields. Two colonies per acre are recommended for almond orchards and about one colony per acre in apple orchards.

Some growers prefer to have the colonies placed alongside the orchard; others want them distributed in small groups within the orchard. Bees also are used regularly by growers of many other crops: blueberries, cantaloupes, cherries, clovers, cucumbers, cranberries, cutflower seed, plums and prunes, vetch, and watermelon.

DISEASE AND PEST CONTROL

Honeybees have diseases and enemies: diseases of the brood; diseases that affect only the adult bees; insect enemies of the adults and of the comb; and other enemies, including toads, lizards, birds, mice, skunks, and bears.

Diseases. American foulbrood, caused by a spore-forming bacterium, *Bacillus larvae,* is the most serious brood disease. It occurs throughout the world wherever bees are kept and affects workers, drones, and queens. The spores are highly resistant to heat and chemicals. A comb containing brood severely infected with this disease has a mottled appearance caused by the mixture of healthy capped brood interspersed with diseased or empty cells formerly occupied by diseased brood. The decayed mass has a typical ropiness when dug into, which is one of its identifying characteristics.

American foulbrood can be spread to healthy colonies by transferring equipment or allowing the bees to feed on honey from infected colonies. Sulfathiazole and Terramycin are widely used to control the disease. Many countries and most states in the U.S. require the destruction by fire of diseased colonies and have apiary inspectors to enforce the regulations.

European foulbrood is caused by a nonsporeforming bacterium, *Streptococcus pluton,* but *Bacillus alvie* and *Acromobacter eurydice* are often associated with *Streptococcus pluton.* This disease is similar in appearance to American foulbrood. In some instances it severely affects the colonies, but they recover so that colony destruction is not necessary. Terramycin can control the disease.

Sacbrood is caused by a virus and is superficially similar to the foulbrood diseases. It can appear and disappear spontaneously but is seldom serious. No chemical control is needed, but if the problem persists the beekeeper usually requeens the colony.

Chalk brood is caused by the fungus *Ascosphaera apis.* The larvae victims of this disease have a chalky, white appearance.

Stonebrood, which affects both brood and adults, is also caused by a fungus, *Aspergillus flavus,* which can usually be isolated from bees that have stonebrood.

Nosema disease

Nosema disease, caused by the protozoan *Nosema apis,* is the most serious disease of adult bees. It is widespread, causes heavy losses in honey production, and severely weakens colonies. The external symptoms of bees with nosema disease are not apparent. The disease is transmitted from adult to adult by ingestion of the spores that soon germinate in the ventriculus, or main, stomach. An infected ventriculus is normally swollen, soft, and grayish white. A degree of control may be obtained by feeding the colony the drug fumagillin.

Acarine disease is caused by the mite *Acarapis woodi* that gets into the tracheae of the bee through its breathing holes or spiracles in its thorax or midsection. Bees affected by this mite are unable to fly, have disjointed wings and distended abdomens. There is presently no good control for this mite. The only U.S. federal law pertaining to bees was passed to prevent the importation of adult bees carrying this mite into the United States. Two other mites, *Varroa jacobsoni* and *Tropilaelops clareae,* are serious problems of Asian beekeepers, but they do not occur in Europe or North America.

There are other minor diseases of adult bees, but they seldom cause serious problems.

Pests. The greater waxmoth, *Galleria mellonella,* is a lepidopterous insect that, in its larval stage, destroys combs. It does not attack adult bees, but may begin destruction of combs of a weak colony long before the bees are gone. It can also destroy stored combs of honey. When the larvae are ready to pupate they often eat out a place to spin their cocoons in the soft wood of the beehive, damaging frames and other hive parts. The best control for this pest is keeping colonies strong. Stored combs are fumigated, kept in a cold room or stacked in such a way that a strong air draft flows around them.

The larvae of the lesser waxmoth, *Achroia grisella,* cause damage to stored combs similar to that of the greater waxmoth. The Mediterranean flour moth larva, *Anagasta kuehniella,* feeds on pollen in the combs and causes some damage. Control for both of these moths is the same as for the greater waxmoth.

The bee louse, *Braula caeca,* is a tiny, wingless member of thc fly family that is occasionally found on bees, but feeds on nectar or honey from the mouthparts of its host. Its larvae burrow in the cappings of honey combs.

Ants sometimes invade hives and disrupt or kill the bees.

Termites can damage or destroy hive parts placed on the soil.

Other insects, such as dragonflies (Odonata), robberflies (Diptera), praying mantises (Orthoptera), ambush bugs (Hemiptera), and certain wasps and yellow jackets (Hymenoptera) are natural enemies of the honeybee.

Predators. Mice frequently enter the hive in winter when the bees are clustered, or they get into stored combs and despoil or damage them by chewing the frames and combs to construct their nest.

Skunks devour large numbers of bees at the hive entrance, usually at night. Fences, traps, and poison are used against them.

Bears eat the honeybees and brood in the hive, and usually destroy it and its contents in the process. In bear country, electric fences and traps are used to protect bee colonies.

Fighting among bees

At times bees become their own deadly enemy. If honey is exposed to them when no flowers are in bloom and the weather is mild, the bees from different colonies will fight over it. Sometimes this fighting, or robbing, becomes intense and spreads from hive to hive in moblike action. If all the bees in one colony are killed, the honey is quickly stolen and carried into other hives. This further intensifies the robbing so that a cluster that was carrying honey into its hive a few minutes earlier is attacked, all of its occupants killed, the honey again stolen and the process repeated. Usually, once robbing becomes intense, only darkness or foul weather will stop it. (S.E.McG.)

Cultivation of specialty crops

COTTON

Cotton, the seed fibre of a variety of plants of the genus *Gossypium,* belonging to the Malvaceae family, is the world's most important nonfood agricultural commodity. This section treats the cultivation of the cotton plant. For detailed information on the processing of cotton fibre and the history of its many uses, see the article INDUSTRIES, TEXTILE.

The cotton plant is normally cultivated as a shrubby annual in temperate climates but can be found as a perennial in treelike plants in tropical climates. The cultivated shrub grows from four to six feet tall (about one to two metres) over a growing period of six to seven months.

Factors influencing cotton crops

Warm and humid climates with sandy soil are the most suitable. Although cotton can be grown between latitudes 30° N and 30° S, yield and fibre quality are considerably influenced by climatic conditions, and best qualities are obtained with high moisture levels resulting from rainfall or irrigation during the growing season and a dry, warm season during the picking period. Rain or strong wind may cause damage to the opened bolls.

Within 80–100 days after planting, the plant develops white blossoms, which change to a reddish colour. The blossoms fall off after a few days and are replaced by small green triangular pods, called bolls, that mature after a period of 55–80 days. During this period the seeds and their attached hairs develop within the boll, which increases considerably in size. The seed hair, or cotton fibre, reaching a maximum length of about two and a half inches (approximately six centimetres) in long fibre varieties, is known as lint. Linters, fibres considerably shorter than the seed hair and more closely connected to the seed, come from a second growth beginning about 10 days after the first seed hairs begin to develop. When ripe, the boll bursts into a white, fluffy ball containing three to five cells, each having seven to 10 seeds embedded in a mass of seed fibres. Two-thirds of the weight of the seed cotton (*i.e.,* the seed with the adhering seed hair) consists of the seeds. To avoid damage by wind or rain the cotton is picked as soon as the bolls open, but since the bolls do not all reach maturity simultaneously, an optimum time is chosen for harvesting by mechanical means. Handpicking, carried out over a period of several days, allows selection of the mature and opened bolls, so that a higher yield is possible. Handpicking also produces considerably cleaner cotton; mechanical harvesters pick the bolls by suction, accumulating loose material, dust, and dirt, and cannot distinguish between good and discoloured cotton. A chemical defoliant is usually applied before mechanical picking to cause the plants to shed their leaves, thus encouraging more uniform ripening of the bolls.

Damage by insects and micro-organisms

Cotton is attacked by several hundred species of insects, including such harmful species as the boll weevil, pink bollworm, cotton leafworm, cotton fleahopper, cotton aphid, rapid plant bug, conchuela, southern green stinkbug, spider mites (red spiders), grasshoppers, thrips, and tarnished plant bugs. Limited control of damage by insect pests can be achieved by proper timing of planting and other cultural practices or by selective breeding of varieties having some resistance to insect damage. Chemical insecticides, which were first introduced in the early 1900s, require careful and selective use because of ecological considerations but appear to be the most effective and efficient means of control.

The boll weevil (*Anthonomus grandis*), the most serious cotton pest in the United States in the early 1900s, was finally controlled by appropriate cultivation methods and by the application of such organic insecticides as chlorinated hydrocarbons and organic phosphates. A species of boll weevil resistant to chlorinated hydrocarbons was recorded in the late 1950s; this species is combatted effectively with a mixture of toxaphene and DDT (dichlorodiphenyltrichloroethane), which has been outlawed in the United States and some other countries, however. The pink bollworm (*Pectinophora gossypiella*), originally reported in India in 1842, has spread throughout the cotton-producing countries, causing average annual crop losses of up to 25 percent in, for example, India, Egypt, China, and Brazil. Controls and quarantines of affected areas have helped limit the spread of the insect, and eradication has been possible in a few relatively small areas with sufficiently strict controls. The bollworm (*Heliothis zea,* also known as the corn earworm) feeds on cotton and many other wild and cultivated plants. Properly timed insecticide application provides fairly effective control.

Cotton plants are subject to diseases caused by various pathogenic fungi, bacteria, and viruses and to damage by nematodes (parasitic worms) and physiological disturbances also classified as diseases. Losses have been estimated as high as 50 percent in some African countries and in Brazil. Because young seedlings are especially sensitive to attack by a complex of disease organisms, treatment of seeds before planting is common. Some varieties have been bred that are resistant to a bacterial disease called angular leaf spot. Soil fumigation moderately succeeded in combatting such fungus diseases as fusarium wilt, verticillium wilt, and Texas root rot, which are restricted to certain conditions of soil, rainfall, and general climate. The breeding of resistant varieties, however, has been more effective. (H.-D.H.W.)

TOBACCO

Tobacco is the common name of the plant *Nicotiana tabacum* and, to a limited extent, *N. rustica* and the cured leaf that is used, usually after aging and processing in various ways, for smoking, chewing, snuffing, and extraction of nicotine. This section deals with the farming of tobacco from cultivation to curing and grading. For information on the history of tobacco cultivation and use and the manufacture of tobacco products, see the article TOBACCO in the *Micropædia*.

Cultivation. Though tobacco is tropical in origin, it is cultivated throughout the world. *N. tabacum* requires a frost-free period of 100 to 130 days from date of transplanting to maturity in the field. *N. rustica,* which is grown to some extent in India and certain Transcaucasian countries, matures in advance of *N. tabacum*.

The prime requisite for successful tobacco culture is a supply of well-developed, healthy seedlings that is available at the proper time for transplanting. Orinoco strains of seed are sown to grow leaf for flue curing. The Pryor group are grown to produce the dark air-cured and fire-cured types. Burley and Maryland strains are seeded for the production of light, air-cured tobaccos. Broadleaf and seed-leaf strains, Havana seed, Cuban, and Sumatra varieties are for the production of cigars. The variety grown for production of Perique resembles the Cuban-like variety used in Puerto Rico. Aromatic varieties are grown for production of this type of leaf and in some degree resemble the Cuban varieties.

Soil for a plant bed should be fertile and of good tilth and drainage; it must be protected from chilling winds and exposed to the sun. The soil is usually partially sterilized by burning, steaming, or using chemicals such as methyl bromide to control diseases, weeds, insects, and nematodes (a class of parasitic worms). In warm regions of the world the small germinating seedlings are produced outdoors in cold frames covered with thin cotton cloth or a thin mulch, such as chopped grass (used in particular in Zimbabwe), straw, or pine needles. Glass or plastic is used in colder regions, and close attention is given to watering and ventilation. The usual rate of seeding—*i.e.,* about one ounce (28 grams) of cleaned seed of high germination to 200 square yards (167 square metres) of seedbed area—can be expected, under favourable conditions, to produce 15,000 to 25,000 plants for transplanting. High-analysis mixtures of commercial fertilizers are usually applied before seeding at the rate of one-half to two pounds per square yard (0.3 to 1 kilogram per square metre) of seedbed area. The soil must be finely pulverized and level so that the seed can be lightly covered with soil by rolling or trampling. Uniform distribution of seeds is important. After eight to 10 weeks the seedlings are four to seven inches (10 to 18 centimetres) in length and are ready for transplanting in the field.

Transplanting machines are used extensively in some areas, but most of the world's tobacco is planted by hand. When the soil is dry, adding water helps a high percentage of transplants to survive. Fumigation of soil prior to transplanting is a common practice in many areas where nematodes are common; the process helps to reduce the damage caused by their parasitic activity.

Soil and fertilizer requirements vary widely with the type of tobacco grown. Well-drained soil with a structure that assures good aeration is desirable. Flue-cured, Maryland, cigar binder, and wrapper types of tobacco are produced on sandy and sandy-loam soil, with a sandy and sandy-clay subsoil where local conditions permit. Burley, dark air-cured, fire-cured, cigar-filler, and cigar-binder types are grown on silt-loam and clay-loam soils, with clay subsoils. The type of tobacco, soil, and climate determine fertilizer requirements. If any of the chemical elements essential for growth are lacking, the tobacco plant develops nutritional deficiency symptoms. Though nitrogen, phosphoric acid, and potash may be applied in the shade cigar-wrapper area of Florida–Georgia, very little fertilizer is used on eastern European fields of aromatic tobacco, where rich soils can make the leaf grow too large and rank to be desirable commercially.

Cigar-wrapper leaf

Soil must be prepared and cultivated to control weeds and promote the early and continuous growth of tobacco. For production of cigar-wrapper leaf, a unique method of

Eastfoto

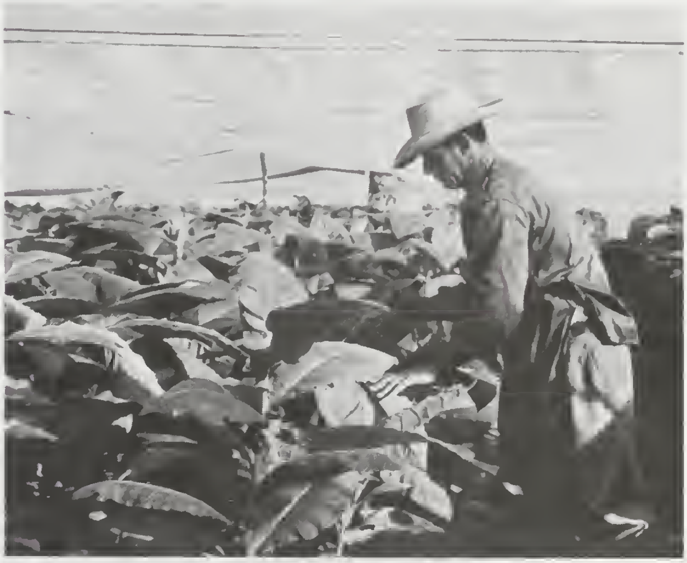


Figure 16: Tobacco cultivated under artificial shading cloth to produce leaves suitable for cigar wrappers, Cuba.

culture is practiced in Cuba and the United States, under artificial cheesecloth shade (see Figure 16). A high moisture content is maintained in soil and air to produce a thin, elastic leaf. In Sumatra and Java under the prevailing conditions of soil and climate, tobacco for cigar-wrapper is produced for one or two years following the clearing of jungle growth. Climatic and soil conditions characterized by a moist atmosphere appear to be associated with the production of acceptable cigar-wrapper tobacco. Cuban leaf for cigar filler is produced on certain soils from special varieties in the prevailing climatic conditions.

Aromatic tobacco culture in Turkey, in such Balkan countries as Bulgaria and Greece, and in certain other areas differs from that of most of the large-leafed tobaccos in that the plants are rarely topped and preferably are grown on soils of low productivity. The most acceptable aromatic leaf is produced in the Mediterranean climate, maturing during dry periods on upland soils.

Spacing of plants in the field varies widely according to the type of tobacco. Flue-cured tobacco rows are four feet (1.2 metres) apart, with plants 20 to 24 inches (50 to 60 centimetres) apart in the row. Burley and cigar tobaccos are three to 3½ feet (1.1 metres) by 15 to 27 inches (38 to 68 centimetres). Dark air-cured and fire-cured tobaccos may be planted on the square with hills 3½ feet apart. Maryland may be planted 32 to 36 inches (81 to 91 centimetres) or closer. Aromatic tobaccos are spaced in rows 15 to 24 inches (38 to 60 centimetres) apart, with three to eight inches (eight to 20 centimetres) between plants in the row. Perique is spaced the widest, with rows five feet (1.5 metres) apart and 36 to 42 inches (91 to 107 centimetres) between plants.

Topping

Large-leaf tobaccos grown in the United States and in several other countries are topped—that is, the terminal growth is removed—when the plant has reached the desired size, usually at or shortly after flowering. The number of leaves remaining varies widely. Dark air-cured and fire-cured tobaccos may have 10 to 16 leaves; Burley, flue-cured, Maryland, and cigar types may have 16 to 20 leaves. After topping, the suckers, or lateral shoots, are removed to increase leaf development, providing increased yields. The work may be done by hand, in which case it must be repeated regularly, or by application of sucker-suppressing chemicals.

Diseases and pests. Common diseases and pests are black root rot, fusarium wilt, tobacco mosaic, bacterial leaf spot, downy mildew or blue mold, black shank, broomrape, and witchweed. These may be controlled by sanitation, crop rotation, the use of sprays and fumigants, and breeding of disease-resistant strains. Resistance to bacterial leaf spot, fusarium wilt, mosaic, black shank, and black root rot have been accomplished by breeding. Some resistant varieties of tobacco in general use have been produced by blending desired characteristics from *N. longiflora, N. debneyi, N. glutinosa,* and others with some strain of *N. tabacum.*

Common insect pests are the green June beetle larvae, cutworms, and flea beetles in the plant bed and hornworms, grasshoppers, flea beetles, cutworms, budworms, and aphids in the field. The cigarette, or tobacco, beetle damages the stored leaf and sometimes the manufactured product. Insect pests are controlled on the growing crop by using sprays and dusts, on the stored product by fumigating and trapping. Biological control often is effective. Fumigation controls nematodes in the field.

Harvest. Tobacco is harvested 70 to 130 days after transplanting by one of two methods: (1) the entire plant is cut and the stalk split or speared and hung on a tobacco stick or lath, or (2) the leaves are removed at intervals as they mature. The leaves of cigar-wrapper and aromatic tobaccos are strung using a needle, and leaves to be flue-cured are looped, using a string tied to a lath or stick that is hung in the curing barn. To prevent breakage and bruising during the handling necessary in curing, it is desirable for the leaf to wilt without sunburning. Tobacco may be left in the field from a few hours to two days to wilt.

Curing. The three common methods of curing are by air, fire, and flue. A fourth method, sun curing, is practiced with aromatic types and to a limited extent with air-cured types. Curing entails four essential steps: wilting, yellowing, colouring, and drying. These involve physical and chemical changes in the leaf and are regulated to develop the desired properties. Air curing (see Figure 17) is accomplished mainly by mechanical ventilation inside buildings. Coke, charcoal, or liquid petroleum gas may be burned to provide heat when conditions warrant. Air curing, which requires from one to two months' time, is used for many tobaccos, including dark air-cured types, cigar, Maryland, and Burley.

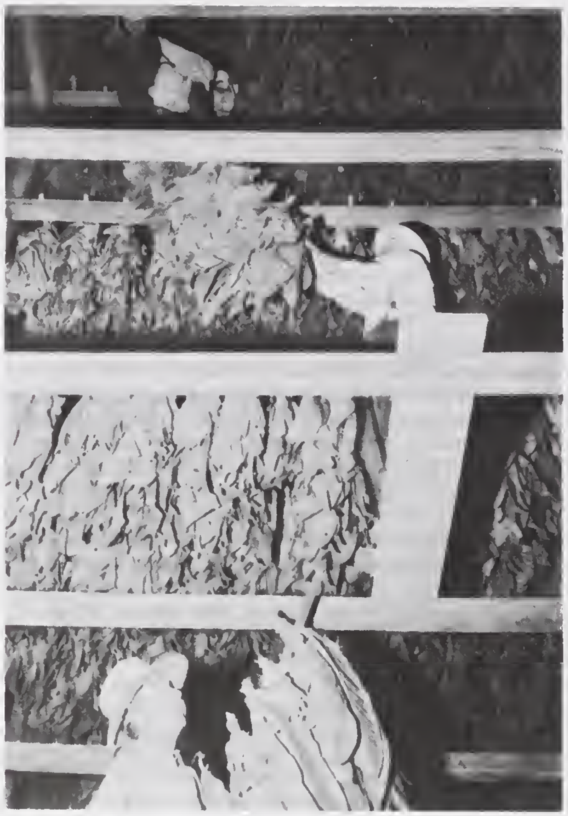


Figure 17: Hanging Maryland tobacco, cured on the stalk, in a curing barn.

By courtesy of the U.S. Department of Agriculture

Air curing, fire curing, and flue curing

The fire-curing process resembles air curing except that open wood fires are kindled on the dirt floor of the curing barn after the tobacco has been hanging for two to six days. The smoke imparts to the tobacco a characteristic aroma of creosote. The firing process may be continuous or intermittent, extending from three weeks to as long as 10 weeks until curing is complete and the leaf has been cured to the desired finish.

The barns for flue curing are small and tightly constructed with ventilators and metal pipes, or flues, extending from furnaces around or under the floor of the barn. Fuels used are wood, coal, oil, and liquid petroleum gas. If oil or gas heaters are used, flues are not needed. Heat is applied carefully, and the leaves are observed closely for changes in their chemical and physical composition. Flue curing requires from four to eight days' time and is used for Virginia, or bright, tobacco. In the process called bulk curing, the leaves are loaded evenly in racks arranged in a curing chamber.

Grading. After curing, the leaf may be piled in bulk to condition for a time before it is prepared for sale. The preparation consists usually of grading the leaf and putting it in a bale or package of convenient size and weight for inspection and removal by the buyer. Except during humid periods, the leaf must be conditioned in moistening cellars or humidified rooms before it can be handled without breakage. Type of leaf and local custom determine the fineness of grading. At its most elaborate, grading may be by position of the leaf on the plant, colour, size, maturity, soundness, and other recognizable qualities; flue-cured tobacco in the United States and Zimbabwe is graded this way, and each grade bulked or baled separately. Much simpler grading is usual in developing countries, where the buyer is as much concerned with the proportions of each grade as with the quality of the entire lot; aromatic tobaccos are an example of this. Most tobaccos entering world trade, except the aromatic, are assembled before sale into bundles, or hands, of 15 to 30 leaves and tied with one leaf wrapped securely around the butts.

Forced aging

Most tobaccos, except aromatic and cigar, are regraded if necessary and usually redried after purchase; then the exact amount of moisture needed for aging is added and the tobacco is securely packed in cases or hogsheads. Exported tobacco is shipped in this form. The trend is for the packing factories to stem the leaf—that is, remove most of the stem leaving the lamina—usually by threshing machines but sometimes by hand, before redrying it. The aging process, particularly with cigar tobaccos, is sometimes hastened by forced fermentation procedures. After purchase, aromatic tobaccos are manipulated; that is, they are factory-graded, baled, and subjected to an elaborate, in-the-bale, fermentation process before going to the ultimate manufacturer. (J.E.McM.)

TEA

The tea plant, a species of evergeen (*Camellia sinensis*), is cultivated for its young leaves and leaf buds, from which the tea beverage is produced. This section treats the cultivation of the tea plant. For information on the processing of tea and the history of its use, see the article BEVERAGE PRODUCTION.

Varieties. The natural habitat of the tea plant is considered to be within the fan-shaped area between the Nāgāland, Manipur, and Lushai hills along the Assam–Myanmar (Burma) frontier in the west; through to China, probably as far as Chekiang province in the east; and from this line south through the hills of Myanmar and Thailand into Vietnam. The three main varieties of the tea plant, China, Assam, and Cambodia, each occur in their most distinct form at the extremes of the fan-shaped area. There are an infinite number of hybrids between the varieties; such crosses can be seen in almost any tea field.

The China variety, a multistemmed bush growing as high as nine feet (2.75 metres), is a hardy plant able to withstand cold winters and has an economic life of at least 100 years. When grown at an altitude near that of Darjeeling and Sri Lanka (Ceylon), it produces teas with valuable flavour during the season's second flush or growth of new shoots.

The Assam variety, a single-stem tree ranging from 20 to 60 feet (six to 18 metres) in height and including several subvarieties, has an economic life of 40 years with regular pruning and plucking. The tea planter recognizes five main subvarieties: the tender light-leaved Assam, the less tender dark-leaved Assam, the hardy Manipuri and Burma types, and the very large-leaved Lushai. In Upper Assam, the dark-leaved Assam plant, when its leaves are highly pubescent, produces very fine quality "golden tip" teas during its second flush. (The Chinese word *pekho*, meaning "white hair" or "down," refers to the "tip" in tea, which is correlated with quality.)

The Cambodia variety, a single-stem tree growing to about 16 feet (five metres) in height, is not cultivated but has been naturally crossed with other varieties.

The mature leaves of the tea plant, differing in form according to variety, range from 1½ to 10 inches (3.8 to 25 centimetres) in length, the smallest being the China variety and the largest the Lushai subvariety. In harvesting, or plucking, the shoot removed usually includes the bud and the two youngest leaves. The weight of 2,000 freshly plucked China bush shoots may be one pound (0.45 kilogram); the same number of Assam shoots may weigh two pounds (0.9 kilogram). Tea leaves may be serrated, bullate, or smooth; stiff or flabby; the leaf pose ranges from erect to pendant; and the degree of pubescence varies widely from plant to plant.

Cultivation. Three considerations in planning a tea estate are climate, soil acidity, and labour availability.

A suitable climate has a minimum annual rainfall of 45 to 50 inches (1,140 to 1,270 millimetres), with proper distribution. If there is a cool season, with average temperatures 20° F (11° C) or more below those of the warm season, the growth rate will decrease and a dormant period will follow, even when the cool season is the wetter one.

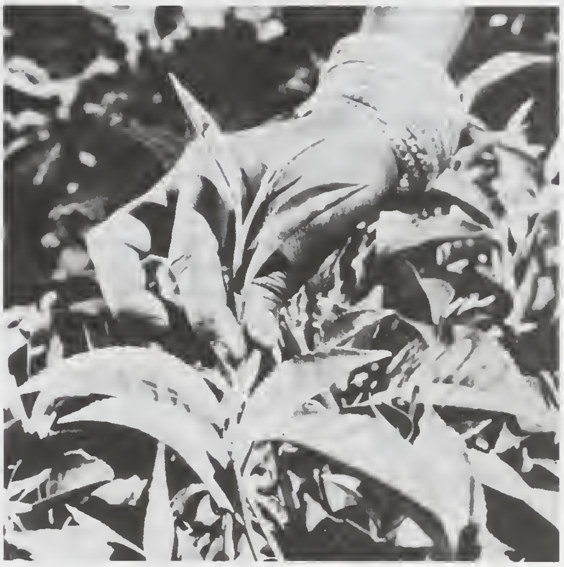


Figure 18: A tea shoot comprising a bud and two leaves being plucked by hand in Sri Lanka.
By courtesy of the Ceylon Tea Bureau, London

Tea soils must be acid; tea cannot be grown in alkaline soils. A desirable pH value is 5.8 to 5.4 or less. A crop of 1,500 pounds of tea per acre (1,650 kilograms per hectare) requires 1.5 to 2 workers per acre (3.7 to 4.9 workers per hectare) to pluck the tea shoots and perform other fieldwork. Mechanical plucking has been tried, but because of its lack of selectivity, cannot replace hand plucking.

The scientific study of tea

Scientific study of tea production began about 1890. Most tea-producing countries maintain scientific research stations to study every aspect of the subject, including seed production, clonal selection (for the propagation of single leaf cuttings), tea nursery management, transplanting, development of the bush and subsequent pruning and plucking, soil management and fertilizer use, and the ultimate replanting of the stand. Although procedures in all countries are related, appropriate details must be determined for each area. Since 1900, advancements in tea cultivation have increased the average yield per acre in Assam from 400 to 1,000 pounds (180 to 450 kilograms), with many estates producing over 1,500 pounds (680 kilograms).

Pests and diseases. The tea plant is subject to attack from at least 150 insect species and 380 fungus diseases. In northeast India, where 125 pests and 190 fungi have been detected, losses from pests and diseases have been

By courtesy of Brooke Bond Oxo Ltd

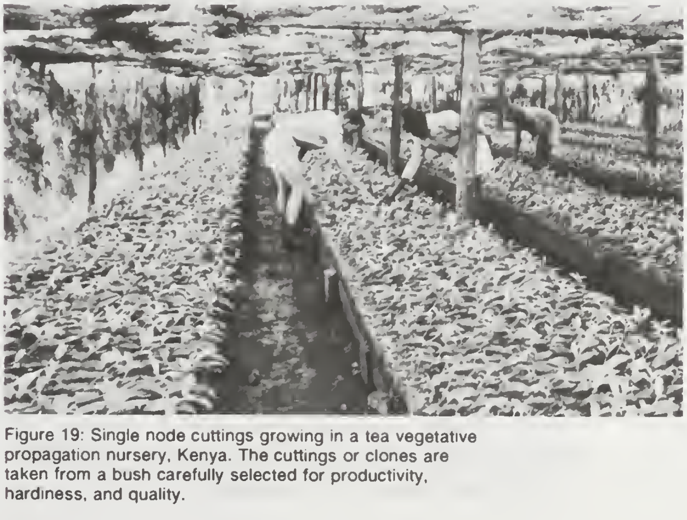


Figure 19: Single node cuttings growing in a tea vegetative propagation nursery, Kenya. The cuttings or clones are taken from a bush carefully selected for productivity, hardiness, and quality.

estimated at 67,000,000 pounds (30,000,000 kilograms) of tea per annum. More than 100 pests and 40 diseases occur in the tea fields of Japan. Sri Lanka, where estates are close together or contiguous, has recorded many blights and suffered serious losses. Africa has little trouble with blights; the tea mosquito (*Helopeltis theivora*) is the only serious pest. The Caucasus, with a climate similar to that of Japan, grows the China variety of plant and has no serious pests or blights.

Blight control has become highly developed. Northeast Indian scientists have issued a list of 40 approved proprietary pesticides. Some of these pesticides cannot be applied during the plucking season; others require that the two subsequent rounds of weekly pluckings be discarded.

(C.R.H.)

BIBLIOGRAPHY

General works. C.B. RICHEY (ed.), *Agricultural Engineers' Handbook* (1961), though dated, is still a useful reference on crop-production equipment, soil and water conservation, and farm buildings. More recent treatments include *Primrose McConnell's The Agricultural Notebook,* 18th ed. edited by R.J. HALLEY and R.J. SOFFE (1988), on all aspects of farming, with special emphasis on practices in the United Kingdom; and GORDON WRIGLEY, *Tropical Agriculture,* 4th ed. (1982), on crop ecology, culture (including different systems), improvement, and protection, with a section on cattle (including water buffalo). The broad range of factors that determine different agricultural systems are discussed in DAVID GRIGG, *An Introduction to Agricultural Geography* (1984). Each issue of *The State of Food and Agriculture* (annual), published by the Food and Agriculture Organization of the United Nations (FAO), in addition to world and regional reviews, includes one or more special studies of longer-term issues. The *Yearbook of Agriculture,* published by the U.S. Dept. of Agriculture, each year focusses on a particular theme—*e.g., Farm Management* (1989), *Agriculture and the Environment* (1991), and *New Crops, New Uses, New Markets* (1992). ALFRED H. KREBS, *Agriculture in Our Lives,* 5th ed. (1984), provides an introduction to agricultural business and natural resources. DOUGLAS M. CONSIDINE (ed.), *Foods and Food Production Encyclopedia* (1982), outlines the growth, harvesting, and processing of a large number of foods from plants and animals.

Agricultural economics. THEODORE W. SCHULTZ, *Transforming Traditional Agriculture* (1964, reprinted 1983), has become a classic on the complex problems involved in the modernization of agriculture. The interrelationships between agriculture and the rest of the economy are detailed in ROBERT D. STEVENS and CATHY L. JABARA, *Agricultural Development Principles* (1988), a beginning text; and ISAAC ARNON, *Modernization of Agriculture in Developing Countries: Resources, Potentials, and Problems,* 2nd ed. (1987). *Ceres* (bimonthly), published by the FAO, addresses agriculture and food production in developing countries. R. ALBERT BERRY and WILLIAM R. CLINE, *Agrarian Structure and Productivity in Developing Countries* (1979), establishes a correlation between small farm size and agricultural productivity in developing countries. D. GALE JOHNSON, *World Agriculture in Disarray,* 2nd ed. (1991), analyzes the relationship between domestic agricultural policies and international trade. Various farming systems are examined by B.L. TURNER II and STEPHEN B. BRUSH (eds.), *Comparative Farming Systems* (1987), case studies of 12 major world systems; ZHORES A. MEDVEDEV, *Soviet Agriculture* (1987), a historical treatment; and KARL-EUGEN WÄDEKIN (ed.), *Communist Agriculture: Farming in the Soviet Union and Eastern Europe* (1990), a look at the Communist approach before its abandonment.

Farm buildings. Illustrated texts include JOHN B. WELLER, *Farm Buildings,* 2 vol. (1965–72), a comprehensive technical book; JAMES S. BOYD and VIRGINIA T. BOYD, *Practical Farm Buildings,* 2nd ed. (1979); JAMES H. WHITAKER, *Agricultural Buildings and Structures* (1979); R.J. LYTLE, *Farm Builder's Handbook, with Added Material for Pole Type Industrial Buildings,* 3rd ed. (1978, reissued 1982); and R.W. BRUNSKILL, *Traditional Farm Buildings of Britain,* 2nd ed. (1987). DAVID SAINSBURY, *Animal Health and Housing* (1967), emphasizes physiologic and sanitary aspects. ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, *Capital and Finance in Agriculture* (1970), discusses the national economic impact of farm buildings in many countries.

Farm management. Principles of farm organization, planning, management, and operation are set forth in DONALD D. OSBURN and KENNETH C. SCHNEEBERGER, *Modern Agricultural Management,* 2nd ed. (1983); JOHN E. KADLEC, *Farm Management* (1985); ROBERT A. LUENING, RICHARD M. KLEMME, and WILLIAM P. MORTENSON, *The Farm Management Handbook,* 7th ed. (1991); and WARREN F. LEE *et al., Agricultural Finance,* 8th ed. (1988). Works specifically on agricultural machinery include HARRIS PEARSON SMITH and LAMBERT HENRY WILKES, *Farm Machinery and Equipment,* 6th ed. (1976); ARCHIE A. STONE and HAROLD E. GULVIN, *Machines for Power Farming,* 3rd ed. (1977); DONNELL HUNT, *Farm Power and Machinery Management,* 8th ed. (1983); and CLAUDE CULPIN, *Farm Machinery,* 12th ed. (1992).

Plant cultivation. Overviews are presented by R.H.M. LANGER and G.D. HILL, *Agricultural Plants,* 2nd ed. (1991), a description of the major varieties and their products; JULES JANICK *et al., Plant Science: An Introduction to World Crops,* 3rd ed. (1981); and J.W. PURSEGLOVE, *Tropical Crops: Dicotyledons,* 2 vol. (1968, reissued in 1 vol., 1974), and *Tropical Crops: Monocotyledons,* 2 vol. (1972, reprinted in 1 vol., 1988). GEORGE W. COX and MICHAEL D. ATKINS, *Agricultural Ecology* (1979), analyzes world grain and vegetable production systems, with an emphasis on the influence of weather. GLENN O. SCHWAB *et al., Soil and Water Conservation Engineering,* 4th ed. (1993), is also instructive.

Works focussing on plant breeding include JOHN MILTON POEHLMAN, *Breeding Field Crops,* 3rd ed. (1987); NEAL F. JENSEN, *Plant Breeding Methodology* (1988); and ABRAHAM BLUM, *Plant Breeding for Stress Environments* (1988). K.K. FRAMJI, B.C. GARG, and S.D.L. LUTHRA, *Irrigation and Drainage in the World,* 3rd ed., rev. and enlarged, 3 vol. (1981–83), outlines the development of irrigation in various countries of the world and describes major projects. B.A. STEWART and D.R. NIELSEN (eds.), *Irrigation of Agricultural Crops* (1990), considers both theoretical and practical aspects. GLENN J. HOFFMAN, TERRY A. HOWELL, and KENNETH H. SOLOMON (eds.), *Management of Farm Irrigation Systems* (1990), is a practical handbook. Factors in soil preparation are analyzed by WILLIAM R. GILL and GLEN E. VANDEN BERG, *Soil Dynamics in Tillage and Traction* (1967); MILTON A. SPRAGUE and GLOVER B. TRIPLETT (eds.), *No-Tillage and Surface-Tillage Agriculture* (1986), a review of these alternatives to traditional plowing; RONALD E. PHILLIPS and SHIRLEY H. PHILLIPS (eds.), *No-Tillage Agriculture: Principles and Practices* (1984); and SAMUEL L. TISDALE *et al., Soil Fertility and Fertilizers,* 5th ed. (1993).

Various cropping systems are described in JOHN VANDERMEER, *The Ecology of Intercropping* (1989); and CHARLES A. FRANCIS (ed.), *Multiple Cropping Systems* (1986). HUBERT MARTIN and DAVID WOODCOCK, *The Scientific Principles of Crop Protection,* 7th ed. (1983), focusses on pest control. Regional variations in farming technique are presented by K.G. BRENGLE, *Principles and Practices of Dryland Farming* (1982); HANS RUTHENBERG *et al., Farming Systems in the Tropics,* 3rd ed. (1980); and L.V. CROWDER and H.R. CHHEDA, *Tropical Grassland Husbandry* (1982). JAMES SHOLTO DOUGLAS, *Advanced Guide to Hydroponics,* new ed. (1985); and HOWARD M. RESH, *Hydroponic Food Production,* 4th ed. (1989), treat this specialized technique.

Weather information is available in *Weekly Weather and Crop Bulletin,* published by the U.S. Dept. of Commerce, Weather Bureau. Studies of agricultural meteorology include RUDOLF GEIGER, *The Climate Near the Ground* (1965; originally published in German, 4th ed., 1961), a classic text; JEN-HU CHANG, *Climate and Agriculture* (1968); ROBERT H. SHAW (ed.), *Ground Level Climatology* (1967); and NORMAN J. ROSENBERG, BLAINE L. BLAD, and SHASHI B. VERMA, *Microclimate,* 2nd ed. (1983). DAVID J. BRIGGS and FRANK M. COURTNEY, *Agriculture and the Environment* (1985), describes temperate agricultural practices and systems and their impact on the environment, with examples from Britain. MERVYN L. RICHARDSON, *Chemistry, Agriculture, and the Environment* (1991), focusses on pesticide and fertilizer pollution from both crop and livestock production. Pollution's effect on agriculture is reported in JAMES J. MACKENZIE and MOHAMED T. EL-ASHRY (eds.), *Air Pollution's Toll on Forests and Crops* (1989).

Animal husbandry. Overviews include JAMES BLAKELY and DAVID H. BADE, *The Science of Animal Husbandry,* 5th ed. (1989); JOHN R. CAMPBELL and JOHN F. LASLEY, *The Science of Animals That Serve Humanity,* 3rd ed. (1985); M.E. ENSMINGER, *Animal Science,* 9th ed. (1991); W.J.A. PAYNE, *An Introduction to Animal Husbandry in the Tropics,* 4th ed. (1990); and CLARENCE E. BUNDY, RONALD V. DIGGINS, and VIRGIL W. CHRISTENSEN, *Livestock and Poultry Production,* 5th ed. (1982). Volumes in the *World Animal Science* series discuss, among other topics, animal health, genetics and breeding, feeds, and the production of several of the livestock species addressed below (including buffalo).

The principles and practices of animal breeding are described by ENOS J. PERRY (ed.), *The Artificial Insemination of Farm Animals,* 4th rev. ed. (1968); EVERETT JAMES WARWICK and JAMES EDWARD LEGATES, *Breeding and Improvement of Farm Animals,* 7th ed. (1979); FREDERICK B. HUTT and BENJAMIN A. RASMUSEN, *Animal Genetics,* 2nd ed. (1982); FRANZ PIRCHNER, *Population Genetics in Animal Breeding,* 2nd ed. (1983; originally published in German, 2nd rev. and expanded ed., 1979); and MALCOLM B. WILLIS, *Dalton's Introduction to Practical Animal Breeding,* 3rd ed. (1991). Textbooks and reference works that are useful sources of information on animal nutrition include OSKAR KELLNER, KRAFT DREPPER, and KLAUS ROHR, *Grundzüge der Fütterungslehre,* 16th completely rev. ed. (1984); LEONARD A. MAYNARD *et al., Animal Nutrition,* 7th ed. (1979); ARTHUR E. CULLISON and ROBERT S. LOWREY, *Feeds and Feeding,* 4th ed. (1987); PETER R. CHEEKE, *Applied Animal Nutrition* (1991); and D.C. CHURCH, *Livestock Feeds and Feeding,* 3rd ed. (1991).

Cereal farming. Y. POMERANZ, *Modern Cereal Science and Technology* (1987), discusses common aspects of cereal grains and their products followed by in-depth descriptions of selected cereals. NEAL C. STOSKOPF, *Cereal Grain Crops* (1985), is also useful for an overview. Various grains and their production are investigated in Y. POMERANZ (ed.), *Wheat: Chemistry and Technology,* 3rd ed. (1988); L.T. EVANS and W.J. PEACOCK (eds.), *Wheat Science, Today and Tomorrow* (1981), a collection of essays on current international wheat research; ROBERT W. JUGENHEIMER, *Corn: Improvement, Seed Production, and Uses* (1976, reprinted 1985); D.H. GRIST, *Rice,* 6th ed. (1986); D.E. BRIGGS, *Barley* (1978); and HUGH DOGGETT, *Sorghum,* 2nd ed. (1988).

Vegetable farming. Production techniques for commercial and private vegetable crops are found in *Knott's Handbook for Vegetable Growers,* 3rd ed. by OSCAR A. LORENZ and DONALD N. MAYNARD (1988); IB LIBNER NONNECKE, *Vegetable Production* (1989); MAS YAMAGUCHI, *World Vegetables: Principles, Production, and Nutritive Values* (1983); and MARK J. BASSETT (ed.), *Breeding Vegetable Crops* (1986).

Fruit farming. Fruit culture is presented in NORMAN FRANKLIN CHILDERS, *Modern Fruit Science,* 9th ed. (1983), a well-illustrated book on deciduous orchard and small fruit culture in the United States from planting to marketing, with extensive bibliographies; STEVEN NAGY and PHILIP E. SHAW, *Tropical and Subtropical Fruits: Composition, Properties, and Uses* (1980); J.A. SAMSON, *Tropical Fruits,* 2nd ed. (1986); JAMES S. SHOEMAKER, *Small Fruit Culture,* 5th ed. (1978), an in-depth culture and literature review of all important small fruits; GENE J. GALLETTA and DAVID G. HIMELRICK (eds.), *Small Fruit Crop Management* (1990); JASPER GUY WOODROOF, *Tree Nuts: Production, Processing, Products,* 2nd ed. (1979), a complete book on temperate and tropical nuts of economic importance; and MICHAEL O'BRIEN, BURTON F. CARGILL, and ROBERT B. FRIDLEY (eds.), *Principles & Practices for Harvesting & Handling Fruits & Nuts* (1983).

Livestock farming. The *Animal Agriculture Series,* whose primary author is M.E. ENSMINGER, consists of several works, each devoted to a specific class of farm animal. Examples include M.E. ENSMINGER, *Horses and Horsemanship,* 7th ed. (1999); and M.E. ENSMINGER and R.C. PERRY, *Beef Cattle Science,* 7th ed. (1997). Other volumes in the series treat dairy cattle, sheep, swine, and poultry, and each is copiously illustrated, covering all aspects of production. Other texts addressing the livestock mentioned in this article include A.L. NEUMANN and KEITH S. LUSBY, *Beef Cattle,* 8th ed. (1986); TILDEN WAYNE PERRY, *Beef Cattle Feeding and Nutrition* (1980); J.L. KRIDER, J.H. CONRAD, and W.E. CARROLL, *Swine Production,* 5th ed. (1982); COLIN T. WHITTEMORE, *Pig Production: The Scientific and Practical Principles* (1980); ELWYN R. MILLER, DUANE E. ULLREY, and AUSTIN J. LEWIS (eds.), *Swine Nutrition* (1991); RON PARKER, *The Sheep Book: A Handbook for the Modern Shepherd* (1983); ALLAN FRASER and JOHN T. STAMP, *Sheep Husbandry and Diseases,* 6th ed. rev. by J.M.M. CUNNINGHAM and JOHN T. STAMP (1987); G.J. TOMES, D.E. ROBERTSON, and R.J. LIGHTFOOT, *Sheep Breeding,* 2nd ed. rev. by WILLIAM HARESIGN (1979); DAVID MACKENZIE, *Goat Husbandry,* 4th ed. rev. and edited by JEAN LAING (1980); and DONALD E. ULMER and ELWOOD M. JUERGENSON, *Approved Practices in Raising and Handling Horses* (1974).

Dairy farming. Dairy production and marketing is treated in WILLIAM M. ETGEN, ROBERT E. JAMES, and PAUL M. REAVES, *Dairy Cattle Feeding and Management,* 7th ed. (1987); G.H. SCHMIDT, L.D. VAN VLECK, and M.F. HUTJENS, *Principles of Dairy Science,* 2nd ed. (1988); and DONALD L. BATH *et al., Dairy Cattle: Principles, Practices, Problems, Profits,* 3rd ed. (1985).

Poultry farming. Domestic fowl production is covered by STUART BANKS, *The Complete Handbook of Poultry-Keeping* (1979); RICHARD E. AUSTIC and MALDEN C. NESHEIM, *Poultry Production,* 13th ed. (1990); and HOMER PATRICK and PHILIP J. SCHAIBLE, *Poultry: Feeds and Nutrition,* 2nd ed. (1980).

Beekeeping. Books that concern the life history of the individual bee and the colony, honey and wax production, diseases of bees, flora that provide nectar and pollen, and economics of beekeeping include JOHN E. ECKERT and FRANK R. SHAW, *Beekeeping* (1960); DADANT & SONS (eds.), *The Hive and the Honey Bee,* extensively rev. (1975); and EVA CRANE, *Bees and Beekeeping: Science, Practice, and World Resources* (1990). A.I. ROOT, *The ABC & XYZ of Bee Culture: An Encyclopedia Pertaining to the Scientific and Practical Culture of Honey Bees,* 40th ed. edited by ROGER A. MORSE and KIM FLOTTUM (1990); and ROGER A. MORSE and TED HOOPER (eds.), *The Illustrated Encyclopedia of Beekeeping* (1985), are useful reference works.

Specialty crops. Coverage of the crops discussed in the article is provided by DAME S. HAMBY (ed.), *The American Cotton Handbook,* 3rd ed., 2 vol. (1965–66), a collection of authoritative contributions on subjects ranging from cotton growing to the final finished fabric; JOHN M. MUNRO, *Cotton,* 2nd ed. (1987); B.C. AKEHURST, *Tobacco,* 2nd ed. (1981); and T. EDEN, *Tea,* 3rd ed. (1976). (H.-D.H.W./C.R.H./Ed.)

Feminism

The belief in the social, economic, and political equality of the sexes, feminism originated largely in the West but is manifested worldwide and is represented by various institutions committed to activity on behalf of women's rights and interests.

Throughout most of Western history, women were confined to the domestic sphere, while public life was reserved for men. In medieval Europe, women were denied the right to own property, to study, or to participate in public life. Even as late as the early 20th century, women in the United States, as in Europe, could neither vote nor hold elective office. Women were prevented from conducting business without a male representative, be it father, brother, husband, legal agent, or even son. Married women could not exercise control over their own children without the permission of their husbands. Moreover, women had little or no access to education and were barred from most professions. In some parts of the world, such restrictions on women continue today.

This article is divided into the following sections:

HISTORY OF FEMINISM

The ancient world. There is scant evidence of early organized protest against such circumscribed status. In the 3rd century BC, Roman women filled the Capitoline Hill and blocked every entrance to the Forum when consul Marcus Porcius Cato resisted attempts to repeal laws limiting women's use of expensive goods. "If they are victorious now, what will they not attempt?" Cato cried. "As soon as they begin to be your equals, they will have become your superiors."

That rebellion proved exceptional, however. For most of recorded history, only isolated voices spoke out against the inferior status of women. In late 14th- and early 15th-century France, the first feminist philosopher, Christine de Pisan, challenged prevailing attitudes toward women with a bold call for female education. Her mantle was taken up by Laura Cereta, a Venetian woman who in 1488 published *Epistolae familiares,* a volume of letters dealing with a panoply of women's complaints, from marital oppression to the frivolity of women's attire. Defenders of the status quo painted women as superficial and inherently immoral, while the emerging feminists proclaimed that women would be the intellectual equals of men if they were given equal access to education.

Early debates on the status of women

The so-called "debate about women" did not reach England until the late 16th century. After a series of satiric pieces mocking women was published in 1589, the first feminist pamphleteer in England, writing as Jane Anger, responded with *Jane Anger, Her Protection for Women.* This volley of opinion continued for almost a century, until another English author, Mary Astell, issued a more reasoned rejoinder in *A Serious Proposal to the Ladies* (1697). Astell suggested that women inclined neither toward motherhood nor a religious vocation should set up secular convents where they might live, study, and teach.

Influence of the Enlightenment. The feminist voices of the Renaissance never coalesced into a coherent movement. This happened only with the Enlightenment, when women began to demand that the new reformist rhetoric about liberty, equality, and natural rights be applied to both sexes. For example, the Declaration of the Rights of Man and of the Citizen, which defined French citizenship after the revolution of 1789, pointedly failed to address the legal status of women.

Olympe de Gouges, a noted playwright, published *Déclaration des droits de la femme et de la citoyenne* ("Declaration of the Rights of Woman and of the Citizen," 1791), declaring woman to be not only man's equal but his partner. The following year Mary Wollstonecraft's *A Vindication of the Rights of Woman,* the seminal English-language feminist work, was published in England. She proposed that women and men be given equal opportunities in education, work, and politics. Women, she wrote, are as naturally rational as men. If they are silly, it is only because society trains them to be irrelevant.

The concept of equal opportunity

The Age of Enlightenment turned into an era of political ferment marked by revolutions in France, Germany, and Italy and the rise of the abolition movement. By the mid-19th century, issues surrounding feminism had added to the tumult of social change, with ideas being exchanged across Europe and North America.

In the first feminist article she signed with her own name, Louise Otto, a German, built on the work of Charles Fourier, a French social theorist, quoting his dictum that "by the position which women hold in a land, you can see whether the air of a state is thick with dirty fog or free and clear." And after Parisian feminists began publishing a daily newspaper entitled *La Voix des femmes* ("The Voice of Women") in 1848, Luise Dittmar, a German writer, followed one year later with her journal, *Soziale Reform.*

The suffrage movement. These debates culminated in the first women's rights convention, held in July 1848 in the small town of Seneca Falls, New York. The idea sprang up during a social gathering of Lucretia Mott, a Quaker preacher and veteran social activist, Martha Wright (Mott's sister), Mary Ann McClintock, Jane Hunt, and Elizabeth Cady Stanton, the wife of an abolitionist and the only non-Quaker in the group. Stanton drafted 11 resolutions for the "Declaration of Sentiments" that guided the Seneca Falls Convention. With Frederick Douglass arguing eloquently on their behalf, all 11 resolutions passed, including the most radical demand—the right to the vote.

Although Seneca Falls was followed by women's rights conventions in other states, the interest quickly faded. Concern in the United States turned to the pending Civil War, while, in Europe, the reformism of the 1840s gave way to the repression of the late 1850s. When the feminist movement rebounded, it became focused on a single issue, woman suffrage, a goal that would dominate international feminism for almost 70 years. Stanton and Susan B. Anthony, a temperance activist, formed the National Woman Suffrage Association in 1869. At first, they based their demand for the vote on the Enlightenment principle of natural law, regularly invoking the concept of inalienable rights granted to all Americans by the Declaration of Independence. By 1900, however, the American passion for such principles as equality had been dampened by a flood of Eastern European immigrants and the growth of urban slums. Suffragist leaders, reflecting that shift in attitude, began appealing for the vote not on the principle of justice or on the common humanity of men and women but on racist and nativist grounds. As early as 1894, in a speech, Carrie Chapman Catt declared that the votes of literate, American-born, middle-class women would balance the votes of foreigners: "[C]ut off the vote of the slums and give to woman . . . the ballot."

This elitist inclination widened the divide between feminists and the masses of American women who lived in those slums or spoke with foreign accents. As a result, working-class women—already more concerned with wages, hours, and protective legislation than with either the vote or issues such as women's property rights—threw themselves into the trade union movement rather than the feminists' ranks. In addition, radical feminists challenged the single-minded focus on suffrage as the sine qua non of women's liberation. Emma Goldman, the nation's leading anarchist, mocked the notion that the ballot could secure equality for women, since it hardly accomplished that for the majority of American men. And Charlotte Perkins Gilman, in *Women and Economics* (1898), insisted that women would not be liberated until they were freed from the "domestic mythology" of home and family that kept them dependent on men.

Alice Paul

Ultimately, mainstream feminist leaders such as Stanton failed to secure the vote for women. It took a different kind of radical, Alice Paul, to reignite the woman suffrage movement in the United States by copying English activists. Like the Americans, British suffragists, led by the National Union of Woman Suffrage Societies, had initially approached their struggle with ladylike lobbying. But in 1903, a dissident faction led by Emmeline Pankhurst began a series of boycotts, bombings, and pickets. Their tactics worked, and, in 1918, the British Parliament extended the vote to women householders, householders' wives, and female university graduates over the age of 30.

Following the British lead, Paul's crusaders organized mass demonstrations and confrontations with the police. In 1920, American feminism claimed a major triumph with the passage of the 19th Amendment to the Constitution.

CONTEMPORARY FEMINISM IN THE WEST

The postsuffrage era. Once the crucial goal of suffrage had been achieved, the feminist movement fractured into a dozen splinter groups. The Women's Joint Congressional Committee fought for legislation to promote education and maternal and infant health care; the League of Women Voters organized voter registration and education drives; and the Women's Trade Union League launched a campaign for protective labour legislation for women.

Each of these groups offered some civic contribution, but none was specifically feminist in nature. Filling the vacuum, the National Woman's Party, led by Paul, proposed an Equal Rights Amendment (ERA) that would ban discrimination based on sex while moving women closer to equality. Many feminists, however, were not looking for strict equality; they were fighting for laws that would benefit women. Paul, however, argued that protective legislation—such as laws mandating maximum eight-hour shifts for female workers—actually limited women's opportunities by imposing costly rules on employers, who would then be inclined to hire fewer women.

Questions abounded. What was feminism—a movement to create full equality, or a movement meant to respond to the needs of women? And if the price of equality was the absence of protection, did women really want equality?

This philosophical dispute was confined to relatively rarefied circles. Throughout the United States, as across Europe, Americans believed that women had achieved their liberation. Women were voting, although in small numbers and almost exactly like their male counterparts. Even Suzanne LaFollette, a radical feminist, concluded in 1926 that women's struggle "is very largely won." Before any flaws in that pronouncement could be probed, the nation—and the world—plunged into the Great Depression. Next, World War II largely obliterated feminist activism on any continent. The war did open employment opportunities for women—from working in factories ("Rosie the Riveter" became an American icon) to playing professional baseball—but these doors of opportunity were largely closed after the war, when women routinely lost their jobs to men discharged from military service. This turn of events angered many women, but few were willing to mount any organized protest. By 1960, the percentage of employed female professionals was down compared with figures for 1930.

The "second wave" of feminism. The women's movement of the 1960s and '70s, the so-called "second wave" of feminism, represented a seemingly abrupt break with the tranquil suburban life pictured in American popular culture following World War II. Yet the roots of the new rebellion were buried in the frustrations of college-educated mothers whose discontent impelled their daughters in a new direction. If first-wave feminists were inspired by the abolition movement, their great-granddaughters were swept into feminism by the civil rights movement, the attendant discussion of principles such as equality and justice, and the revolutionary ferment caused by protests against the Vietnam War.

Women's concerns were on President John F. Kennedy's agenda even before this public discussion began. In 1961 he created the President's Commission on the Status of Women and appointed Eleanor Roosevelt to lead it. Its report, issued in 1963, firmly supported preparing women for motherhood. But it also documented a national pattern of employment discrimination, unequal pay, legal inequality, and meagre support services for working women that needed to be corrected through legislative guarantees of equal pay for equal work, equal job opportunities, and expanded child-care services. The Equal Pay Act of 1963 offered the first guarantee, and the Civil Rights Act of 1964 was amended to bar employers from discriminating on the basis of sex. Yet some deemed these measures insufficient in a country where classified advertisements still segregated job openings by sex, where state laws restricted women's access to contraception, or where incidences of rape and domestic violence remained undisclosed.

Renewed activism

Dissension and debate. Mainstream groups such as the National Organization for Women (NOW) launched a campaign for legal equity, while ad hoc groups staged sit-ins and marches for any number of reasons—from assailing college curricula that lacked female authors to promoting the use of the word *Ms.* as a neutral form of address. Health collectives and rape crisis centres were established. Children's books were rewritten to obviate sexual stereotypes. Women's studies departments were founded at colleges and universities. Protective labour laws were overturned. Employers found to have discriminated against female workers were required to compensate with back pay. Excluded from many industries for decades, women began finding jobs as pilots and construction workers, bankers and bus drivers.

Unlike the first wave, second-wave feminism provoked theoretical discussion about the origins of women's oppression, the nature of gender, and the role of the family. Kate Millett's *Sexual Politics* made the best-seller list in 1970, and in it she broadened the term *politics* to include all "power-structured relationships" and posited that the personal was actually political. Shulamith Firestone, a founder of the New York Radical Feminists, published *The Dialectic of Sex* in the same year, insisting that love disadvantaged women by shackling them to men. One year later, Germaine Greer, an Australian living in London, published *The Female Eunuch,* in which she argued that the sexual repression of women cuts them off from the creative energy they need to be independent and self-fulfilled.

Competing ideologies

Any attempt to create a coherent, all-encompassing feminist ideology was doomed. Even the term *liberation* could mean different things to different people. Feminism became a river of competing eddies and currents. "Anarcho-feminists," who found a larger audience in Europe than in the United States, said that women could not be liberated without dismantling such institutions as the family, private property, and state power. Individualist feminists, calling on libertarian principles of minimal government, broke with most other feminists over the issue of turning to government for solutions to women's problems. "Amazon feminists" celebrated the mythical female heroine and advocated liberation through physical strength. And separatist feminists, including many lesbian feminists, preached that women could not possibly liberate themselves without at least a period of separation from men.

Ultimately, three major streams of thought surfaced. The first was liberal, or mainstream, feminism, which focused on pragmatic change at institutional and governmental levels. Its goal was to integrate women into the power structure and to give women equal access to positions men had traditionally dominated. While aiming for strict equality (to be evidenced by an equal number of women and men in positions of power, or an equal amount of money spent on male and female student athletes), mainstream feminism nonetheless supported protective legislation such as special workplace benefits for mothers.

In contrast to this pragmatic approach, radical feminism aimed to reshape society and restructure its institutions, which were seen as inherently patriarchal. Providing the core theory for modern feminism, radicals argued that women's subservient role in society was too closely woven into the social fabric to be unraveled without a revolutionary revamping of society itself. They strove to supplant hierarchical and traditional power relationships they saw as reflecting a male bias.

Finally, cultural or "difference" feminism, the last of the three currents, rejected the notion that men and women are intrinsically the same and celebrated women's differences. Inherent in its message was a critique of mainstream feminism's attempt to enter traditionally male spheres.

The race factor. Like first-wave feminism, the second wave was largely defined and led by the educated middle-class white women who built the movement primarily around their own concerns. This created an ambivalent, if not contentious, relationship with women of other classes and races. Many black women had difficulty seeing white women as their feminist sisters; in the eyes of many African Americans, after all, white women were as much the oppressor as white men. "How relevant are the truths, the experiences, the findings of White women to Black women?" asked Toni Cade Bambara in *The Black Woman: An Anthology* (1970). "I don't know that our priorities are the same, that our concerns and methods are the same." Yet during the first conference of the National Black Feminist Organization, held in New York City in 1973, activists acknowledged that many of the goals central to the mainstream feminist movement—day care, abortion, maternity leave, violence—were critical to African American women as well. On specific issues, then, African American feminists and white feminists built an effective working relationship.

GLOBALIZATION OF FEMINISM

Twentieth-century European and American feminism eventually reached into Asia, Africa, and Latin America. As this happened, women in developed countries, especially intellectuals, were horrified to discover that women in some countries were required to wear veils in public or to endure forced marriage, female infanticide, widow burning, or clitoridectomy. Many Western feminists soon perceived themselves as saviours of Third World women, little realizing that their perceptions of and solutions to social problems were often at odds with the real lives and concerns of women in these regions. In many parts of Africa, for example, the status of women had begun to erode significantly only with the arrival of European colonialism. In those regions, then, the notion that patriarchy was the chief problem—rather than European imperialism—seemed absurd.

Questioning Western views of feminism

The conflicts between women in developed and developing nations have played out most vividly at international conferences. After the 1980 World Conference of the United Nations Decade for Women: Equality, Development and Peace in Copenhagen, women from less-developed nations complained that the veil and female genital surgery had been chosen as conference priorities without consulting the women most concerned. It seemed that their counterparts in the West were not listening to them. During the 1994 International Conference on Population and Development in Cairo, women from the Third World protested outside because they believed the agenda had been hijacked by Europeans and Americans. The protesters had expected to talk about ways that underdevelopment was holding women back. Instead, conference organizers chose to focus on contraception and abortion. "[Third World women] noted that they could not very well worry about other matters when their children were dying from thirst, hunger or war," wrote Azizah al-Hibri, a scholar of Muslim women's rights. "The conference instead centered around reducing thc number of Third World babies in order to preserve thc earth's resources, despite (or is it 'because of') the fact that the First World consumes much of these resources."

Woman and her child during a demonstration, demanding women's rights and the banning of degrading representations of women, held on International Women's Day, March 8, 2001, in New Delhi, India.
Arko Datta—AFP/Corbis

The influence of feminism

Still, around the world, women are advancing their interests, although often in fits and starts. Feminism has been derailed in countries such as Afghanistan, where the staunchly reactionary and antifeminist Taliban even banned the education of girls. Elsewhere, feminism has achieved significant gains for women, with the eradication of female genital surgery in many African countries and government efforts to end widow burning in India. More generally, and especially in the West, feminism has influenced every aspect of contemporary life, communication, and debate, from the heightened concern over sexist language to the rise of academic fields such as women's studies and ecofeminism. Sports, divorce laws, sexual mores, organized religion—all have been affected, in many parts of the world, by feminism.

Yet questions remain: How will Western feminism deal with the dissension in its ranks, from women who believe the movement has gone too far and grown too radical? How uniform and successful can feminism be at the global level? Can the problems confronting women in the mountains of Pakistan or the deserts of the Middle East be addressed in isolation, or must such issues be pursued through international forums? Given the unique economic, political, and cultural situations that vary from country to country, the answer to these questions may look quite different in Nairobi than in New York.

BIBLIOGRAPHY. ALICE S. ROSSI (ed.), *The Feminist Papers: From Adams to de Beauvoir* (1973, reprinted 1988), collects the key works of feminism. ROSEMARIE PUTNAM TONG, *Feminist Thought: A More Comprehensive Introduction,* 2nd ed. (1998), provides a comprehensive map of 20th-century feminist thinking. Perspectives from around the world are portrayed in EUGENIA C. DELAMOTTE, NATANIA MEEKER, and JEAN F. O'BARR (eds.), *Women Imagine Change: A Global Anthology of Women's Resistance from 600 B.C.E. to Present* (1997), a representation of women from 30 countries; and MARLENE LEGATES, *Making Waves: A History of Feminism in Western Society* (1996), a comprehensive survey of feminism in Europe, the United States, Canada, and Latin America dating from early Christian times to the present. Texts focusing on feminism in the United States include JANE RENDALL, *The Origins of Modern Feminism: Women in Britain, France, and the United States, 1780–1860* (1984, reissued 1990); and ELEANOR FLEXNER and ELLEN FITZPATRICK, *Century of Struggle: The Woman's Rights Movement in the United States,* enlarged ed. (1996). Questions of class and culture are treated in M. JACQUI ALEXANDER and CHANDRA TALPADE MOHANTY (eds.), *Feminist Genealogies, Colonial Legacies, Democratic Futures* (1997); while UMA NARAYAN, *Dislocating Cultures: Identities, Traditions, and Third-World Feminism* (1997), identifies some misrepresentations of feminist agendas in Third World cultures. KAREN OFFEN, *European Feminisms, 1700–1950: A Political History* (2000), looks at the development of European feminism through the mid-20th century. (E.C.B.)

Ferns and Other Lower Vascular Plants

Vascular plants are those that possess a specialized conducting system for the transport of water, minerals, and food materials, as opposed to the more primitive bryophytes—mosses and liverworts—which lack such a system. They include both the seed plants—angiosperms and gymnosperms, the dominant plants on Earth today—and plants that reproduce by spores—the ferns and other so-called lower vascular plants. The lower vascular plants include the ferns, club mosses, spike mosses, quillworts, horsetails, and whisk ferns. Once considered of the same evolutionary line, these plants were formerly placed in a single group, Pteridophyta, and were known as the ferns and fern allies. Although modern studies have shown that the plants are not in fact related, these terms are still used in discussion of the lower vascular plants.

The pteridophytes represent the oldest of land plants. In their early evolution (during the Devonian and Carboniferous periods, 408 to 286 million years ago), there were many forms that are now extinct. The sphenophytes, for example, were once a large and diverse group of herbs, shrubs, vines, and trees but are now limited to only 15 species of horsetails; the woody lycophytes (club mosses) are entirely gone, leaving only a faint trail in their reduced modern representatives. Much of the fossil fern foliage of the Carboniferous Period is of the uncharacteristic seed ferns, which are the probable antecedents of the flowering plants. Modern ferns represent an explosion of evolution in Cretaceous times (144 to 66.4 million years ago).

The pteridophytes are not an economically important group. Though they are used locally by peoples around the world for medicines and food, their greatest value today is in horticulture (ferns). Their remains, however, provide the bulk of the world's coal beds, and their relatively simple structure and life cycle make them extremely valuable to researchers in understanding the overall picture of plant structure and evolution.

A discussion of all types of plants is found in the article PLANTS. For a discussion of the other types of vascular plants, see GYMNOSPERMS and ANGIOSPERMS. For a discussion of the nonvascular plants, see BRYOPHYTES.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 313, and the *Index.*

The article is divided into the following sections:

General form and function of lower vascular plants

VASCULAR SYSTEM

The conduction system of vascular plants includes the xylem, composed largely of tracheids (tubular cells) in the lower vascular plants and gymnosperms and vessels in angiosperms, for conduction of water and minerals; and the phloem (sieve cells) for conduction of food materials. These vascular tissues are arranged in different patterns in different plant groups and in different parts of the plant.

Structure of the vascular cylinder

The vascular cylinder of a stem or root is called the stele. The simplest and apparently most primitive type of stele is the protostele, in which the xylem is in the centre of the stem, surrounded by a narrow band of phloem. It in turn is bounded by a pericycle of one or two cell layers and a single cell layer of endodermis. The pericycle is generally the layer giving rise to the branches in roots, and the endodermis seems to regulate the flow of water and dissolved substances from the surrounding cortex. More common in fern stems are siphonosteles, having a pith in the centre with the vascular tissue forming a cylinder around it. Where a fern leaf is attached to a stem, a part of the vascular tissue of the stem goes into it (a leaf trace), making a slight gap, filled by parenchyma cells (generalized plant cells), in the vascular cylinder. If the leaves are distant and the stem long and creeping, a single gap will be seen in cross section; if leaves are close together or numerous, the gaps overlap, causing the cylinder to appear in cross section as a ring of disconnected round or elongate bars of vascular tissue.

Generally in pteridophytes, when the young organs mature, no further growth in diameter takes place. In several extinct groups a special ring of cells, the cambium, produced additional xylem to the inside and phloem cells to the outside (secondary growth as opposed to primary growth achieved by apical activity of the stem and root), resulting in increased diameter and a truly woody plant. This is common in many seed plants today, but in the extant pteridophytes only two genera (*Botrychium* and *Isoetes*) show a slight vestige of secondary growth. Even in today's tree ferns (*Cyathea, Dicksonia, Cibotium*), with trunks up to 25 metres (80 feet) tall, the tissues are entirely the result of growth from the stem apex. Their strength is derived not from woody growth in diameter but by strengthening tissues surrounding the vascular bundles and in some cases by a mantle of roots.

CELL TYPES

Cells of the vascular system. The cells of the vascular strands in pteridophytes are mainly tracheids, sieve cells, parenchyma, and endodermal cells. The tracheids, which comprise the xylem, or water-conducting tissue, are normally long, narrow, and attenuated at the tips. Their secondary walls display ladderlike (scalariform) thickenings. The largest tracheids are several centimetres long, but most are much smaller. Vessel cells, which have evolved in several lines of fern evolution and are the principal water-conducting cell type of flowering plants, are modified tracheids in which the end walls have lost their primary membranes, thus providing direct, unimpeded connections for water transport between the cells. Vessels, longitudinal channels composed of linear series of such perforated cells, have been reported from such diverse ferns as waterclover (*Marsilea*) and bracken (*Pteridium*).

The phloem is composed mainly of sieve cells—narrow, elongated units that differ from the tracheids in having persistent protoplasts and nuclei (*i.e.,* they are still alive at functional maturity) and in lacking secondary walls with elaborate pitting. Sieve cells usually display more or less distinguishable sievelike areas, through which, presumably, organic foods pass in their travels through the stem and other plant organs. There are various arrangements of xylem and phloem, but usually a single strand composed of both is surrounded by parenchyma cells, the pericycle (a thin zone of living cells just within the endodermis), and an outer layer of cells with specialized walls, the endodermis. Endodermal cells in young stems are provided with special strips of secondary wall material known as Casparian strips on their radial walls (*i.e.,* on all the cell walls except the two that face toward the stem axis and the stem surface). As the stems age, however, there is a tendency for the endodermal cells to become thick-walled around the entire circumference.

Endodermal cells

Other cells. The pith is made up of parenchyma cells as a rule, but, in some fern genera, scattered tracheid-like cells are found as well. The cells of pteridophyte stems differ from those of many seed plants in lacking collenchyma (modified parenchyma cells with expanded primary walls) and true stone cells. Latex-producing cells in lower vascular plants are rare.

ROOTS

Taproots are unknown in lower vascular plants. All pteridophyte roots are referred to as adventitious, in the sense that they arise at points along the stem. In internal structure, the roots are generally regarded as being much less diverse than the stems. They are protostelic, lacking pith and gaps, and they grow from one or more apical initials (cells that divide to produce all the cells and tissues of an organ), producing a root cap outwardly and the permanent tissues of the root inwardly. They entirely lack secondary growth (continued growth in thickness).

The surface cells of the epidermis produce root hairs near the root apex. These cells are generally thin-walled, in contrast to the cells of the cortex, lying below the surface, which ultimately may become very thick-walled. The root hairs have fundamental importance in absorption of water and nutrients and in attachment of the plant to the soil or other growing surface. The endodermis of the root is well marked, and Casparian strips are present, as in the stem. There is also a tendency for the endodermis in older parts of the roots to become thick-walled and hardened (sclerified).

The production and development of xylem tissue in the steles of most pteridophyte roots is diarch; that is, the first matured xylem appears along two lines at the outer periphery of the xylem strand. The xylem is surrounded by phloem, and the branch roots arise from the pericycle.

LEAVES

Stem appendages known as leaves take various forms that evolved independently in different groups of lower vascular plants. The simplest are scalelike emergences, or enations, that are not served by vascular tissue (*i.e.,* they have no veins), found in some extinct groups and in modern whisk ferns (*Psilotum*). The lycophytes have scalelike, needlelike, or awl-shaped "microphylls" with a single, unbranched vein. The sphenophytes have "sphenophylls"—scalelike leaves with a single vein in the modern *Equisetum* or wedge-shaped leaves with a dichotomously forking vein system in many of the fossil forms. These leaf forms are all so simple that the vascular connection with the stem stele does not affect the stele configuration and causes no leaf gap. On the other hand, the complex leaves of ferns (pteridophylls, or megaphylls) probably evolved from a branching stem system and affect the stele by drawing out enough vascular tissue to cause a leaf gap.

REPRODUCTION

The life cycle of pteridophytes exhibits an alternation of generations between gametophytes and sporophytes. The gametophytes are sexual plants producing eggs or sperm or both, and the sporophytes are asexual plants producing spores that are capable of producing new gametophytes. The sporophyte of lower vascular plants, in contrast to that of mosses and liverworts, is obviously the dominant generation. Unlike seed plants, which also have dominant sporophytes, pteridophytes reproduce not by forming seeds but by producing spores—minute single cells covered by a protective wall and readily carried by the wind. The life cycle of these plants is referred to as pteridophytic, or fernlike, as opposed to spermatophytic (seed-plant-like).

Sporophytes

The plant begins life as a spore. The germinating spore grows into a small gametophyte, or prothallium, usually only 0.3 to 1 centimetre (0.2 to 0.4 inch) long or broad, bearing rhizoids (hairlike structures for water and mineral absorption and attachment to the soil). Gametophytes may be green, occurring on the soil surface, or colourless, occurring under the soil (usually saprophytically, with the aid of a mycorrhizal fungus). Sex organs, called antheridia and archegonia, produce sperm and eggs, respectively. The sperm require water in which to swim to the egg for fertilization. The fertilized egg, or zygote, contains one set of chromosomes from each of the two sex cells. The zygote then divides, developing into an embryo, which in turn develops the first leaf, root, and stem apex. The resulting plant, the sporophyte, is the characteristic plant that is normally seen. At maturity, sporangia (spore cases) are produced; in them the spore mother cells divide by a special nuclear division, meiosis, in which the chromosome number is reduced to a single set for each of four resulting spores.

In most pteridophytes all the spores of each plant are alike, and the plant is said to be homosporous. A few groups (the lycophytes *Selaginella* and *Isoetes* and, among the ferns, the water-fern families Marsileaceae, Salviniaceae, and Azollaceae) are heterosporous, forming two types of spores. These plants have two kinds of sporangia, one producing a few large megaspores (holding food reserves for the early development of the embryo) and the other producing many small microspores. The microspore divides to form a reduced gametophyte, merely a jacket of cells and a few sperm cells; the megaspore divides to form a mass of tissue and archegonia, each enclosing an egg.

The life cycle of the lower vascular plants is basically the same as that of seed plants. The main difference is that in seed plants the new young sporophyte (embryo) is kept within a structure (seed) on the parent plant before dispersal and perhaps a resting stage, whereas in lower vascular plants dispersal and resting take place in the spore before the embryo is formed.

Ferns

Ferns are plants belonging to the vascular plant division Filicophyta, having leaves with branching vein systems and the young leaves usually unrolling from a tight fiddlehead, or crosier. The number of fern species is usually placed at approximately 12,000, but estimates range from as low as 9,000 to as high as 15,000, the number varying because certain groups are as yet poorly studied and new species are still being found in unexplored tropical areas. The ferns constitute an ancient division of vascular plants, some of them as old as the Carboniferous Period (beginning 360 million years ago) and perhaps older. Their type of life cycle, dependent upon spores for dispersal, long preceded the seed-plant life cycle.

GENERAL FEATURES

Diversity of fern types

Size range and habitat. The ferns are extremely diverse in habitat, form, and reproductive methods. In size alone they range from minute filmy plants only 2 to 3 millimetres (0.08 to 0.12 inch) tall to huge tree ferns 10 to 25 metres (30 to 80 feet) in height. Some are twining vines; others float on the surface of ponds. The majority of ferns inhabit warm, damp areas of the Earth. Growing profusely in tropical areas, ferns diminish in number with increasingly higher latitudes and decreasing supplies of moisture. Few are found in dry, cold places.

Some ferns play a role in ecological succession, growing from the crevices of bare rock exposures and in open bogs and marshes prior to the advent of forest vegetation. The best-known fern genus over much of the world, *Pteridium,* the bracken, is characteristically found in old fields, where in most places it is ultimately succeeded by woody vegetation.

Distribution and abundance. Geographically, ferns are most abundant in the tropics. Arctic and Antarctic regions possess few species. On the other hand, a small, tropical country such as Costa Rica may have more than 900 species of ferns—more than twice as many as are found in all of North America north of Mexico. The finest display of fern diversity is seen in the tropical rain forests, where in only a few acres more than 100 species may be encountered, some of which may constitute a dominant element of the vegetation. Also, many of the species grow as epiphytes upon the trunks and branches of trees. A number of families are almost exclusively tropical (*e.g.,* Marattiaceae, Gleicheniaceae, Grammitidaceae, Schizaeaceae, Cyatheaceae, Blechnaceae, Davalliaceae). Most of the other families occur in both the tropics and the temperate zones. Only certain genera are primarily temperate and Arctic (*e.g., Athyrium, Cystopteris, Dryopteris, Polystichum*), and even these tend to extend into the tropics, being found at high elevations on mountain ranges and volcanoes.

Ferns as weeds are uncommon, although a few occur. The most notorious is the bracken (*Pteridium*), which spreads quickly by its underground, ropelike rhizome, rapidly invading abandoned fields and pastures in both temperate and tropical regions. One species of water spangles (*Salvinia auriculata*) became a major pest in India, blocking irrigation ditches and rice paddies. Another species (*S. molesta*) within three years covered 520 square kilometres (200 square miles) of the artificial Lake Kariba in southern Africa, cutting off light and oxygen and thus killing other plant life and fish. Some fern species have been introduced into tropical or subtropical areas (*e.g.,* southern Florida and Hawaii) and in some cases have become naturalized and have spread into the native forest. Examples include the giant polypody (*Microsorium scolopendrium*), climbing ferns (*Lygodium japonicum* and *L. microphyllum*), green cliff brake (*Pellaea viridis*), silver fern (*Pityrogramma calomelanos*), Japanese holly fern (*Cyrtomium falcatum*), rosy maidenhair (*Adiantum hispidulum*), Cretan brake (*Pteris cretica*), and ladder brake (*P. vittata*). Two Old World species (*Thelypteris dentata* and *T. torresiana*) were introduced into tropical America beginning about 1930 and now are among the most common species even in some remote areas.

Because of their ability to disperse by spores and their capacity to produce both sex organs on the same gametophyte and thus to self-fertilize, it would seem logical to assume that the ferns possess higher powers of long-distance dispersal and establishment than do seed plants. This conclusion tends to be supported by the fact that remote disjunctions—separated growing regions—in range are frequent among ferns. There are interisland and intercontinental disjunctions, east and west, as well as wide north-south disjunctions including species found in the Northern and Southern hemispheres that skip the tropics. Some disjunctions seem to follow the pattern of prevailing winds; the main centre of distribution of a species often may lead to downwind groups consisting of one or a few small populations sometimes hundreds or thousands of miles away. Examples of species exhibiting west-to-east transcontinental disjunctions in North America are Wright's cliff brake (*Pellaea wrightiana*), mountain holly fern (*Polystichum scopulinum*), and forked spleenwort (*Asplenium septentrionale*); all of these ferns are well known in the western United States, and they exist as tiny populations in the mountains of the eastern states as well. Some species are disjunct between continents, such as between New Zealand and South America (*Blechnum penna-marina* and *Hypolepis rugosula*) or South Africa and Australia and New Zealand (*Todea barbara*). Some disjunct patterns, such as similar plants growing in Asia and in eastern North America, are not the result of long-range dispersal but rather are the remnants of an ancient continuous flora, the intervening areas having been changed over time.

Disjunctions in range

Importance to humans. As a group of plants, ferns are not of great economic value. Many different species have been used as a minor food source and for medicine in various parts of the world. Edible fern crosiers (young leaves with coiled, hook-shaped tips) are popular in some areas. The ostrich fern (*Matteuccia*) of northeastern North America is frequently eaten, apparently with no ill effect, but the two ferns most commonly consumed in East Asia (*Osmunda* and *Pteridium*) have been shown to be strongly carcinogenic. The minute, aquatic mosquito fern (*Azolla*) has become a valuable plant, especially in Southeast Asia; a blue-green alga (*Anabaena azollae*) is always found in pockets on the leaves of *Azolla* and helps convert nitrogen to a form usable by other plants, thus greatly increasing the productivity of rice paddies where the fern occurs. The greatest economic value of ferns has been in horticulture, with large nurseries supplying millions of plants annually for both indoor decoration and outdoor gardens and landscaping.

A major value of ferns is in biological research, for they have retained a primitive life cycle involving two separate and more or less independent generations, or growth phases, the plants of which are wholly different in many respects.

NATURAL HISTORY

Life cycle. The typical fern, a sporophyte, consists of stem, leaf, and root; it produces spores; and its cells each have two sets of chromosomes, one set from the egg and one from the sperm. The sporophyte of most ferns is perennial (it lives for several years) and reproduces vegetatively by branching of the rootlike underground stem, or rhizome, often forming large, genetically uniform colonies, or clones. A few ferns propagate by root proliferations, and some, especially in the wet tropics, reproduce by leaf proliferations.

The spores are haploid; that is, they have one set of chromosomes. They are produced in specialized organs—the spore cases, or sporangia—on the fern leaves (sporophylls). Once released, the spores are carried by wind currents, and a small percentage of them fall in appropriate germination sites to form the sexual plants, or gametophytes. In ferns the gametophytes are commonly referred to as prothallia,

Spore dispersal

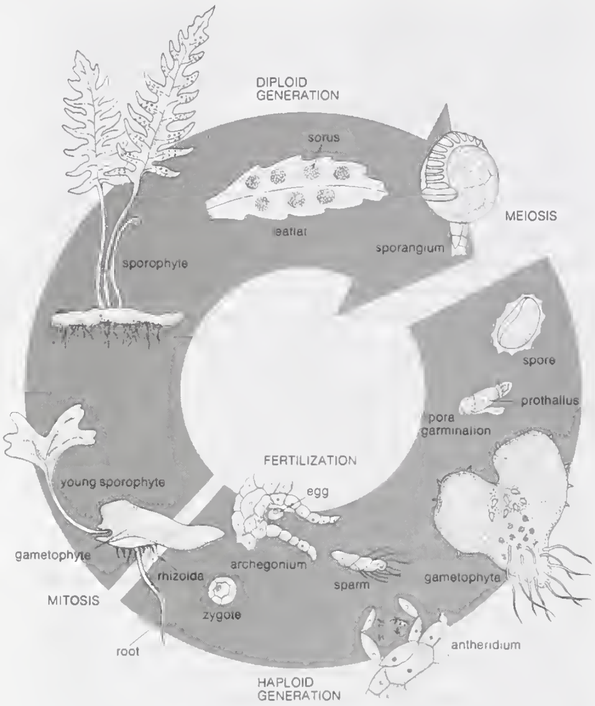


Figure 1: Fern life cycle.
Drawing by M. Pahl

and they are best known to biologists as laboratory objects in artificial culture. They are rarely observed in nature without arduous searching, and the gametophyte stage of the majority of fern species has never been seen in the wild.

The prothallia are tiny—usually less than 8 millimetres (0.3 inch) long—and kidney-shaped in the majority of species. They grow only until the new sporophyte is formed by fertilization; then they wither and die. The process of fertilization is accomplished by sperm and eggs produced upon the same or different gametophytes, and both the fertilized egg (zygote) and the resultant embryo are held within the tissues of the prothallium until the embryo grows out as an independent plant.

Ecology. Ecologically, the ferns are mostly plants of shaded, damp forests of both temperate and tropical zones. Some fern species grow equally well on soil and upon rocks; others are confined strictly to rocky habitats, where they occur in fissures and crevices of cliff faces, boulders, and taluses. Acidic rocks such as granites, sandstones, and quartzites are associated with characteristic fern species different from those of alkaline rocks such as calcites and dolomites. A few species appear to be confined to serpentine and related rocks. In the tropics as many as two-thirds of the ferns of an area may grow as epiphytes on the shaded lower trunks and branches or in the crowns of trees. A few so-called epiphytic ferns are actually climbers that originate upon the ground and grow up tree trunks. In these the lower leaves (bathyphylls) are usually sterile and often different in form from those at the higher levels (acrophylls), which are entirely or partly fertile in that they bear sporangia over their surfaces.

Both epipetric (growing on rocks) and epiphytic ferns may show structural adaptations to dry habitats similar to those of desert plants. These adaptive features include such specializations as hard tissues and thick texture; the surface cells, or epidermis, may be provided with a very thick cuticle (a waxy layer); and abundant hairs or scales may be found on the leaf and stem surfaces. Terrestrial ferns, growing on the ground, may also possess such modifications, especially those that grow in salt marshes (*e.g.,* leather ferns, *Acrostichum*) and open, fully exposed prairies and savannas (*e.g.,* bracken, *Pteridium;* lip ferns, *Cheilanthes;* brakes, *Pteris*).

Ferns that grow in the open are often referred to as sun ferns (*e.g., Gleichenia*) and do not (at least as mature plants) require shade, as do most ferns. Water ferns—waterclovers (*Marsilea*), water spangles (*Salvinia*), and mosquito ferns (*Azolla*)—surprisingly are very commonly inhabitants of dry regions. They appear only after rains, however, and their growth and life cycles are accomplished rapidly, probably as an adaptation to the need for making quick use of water. These ferns have two types of spores that essentially lack the vegetative phase of other ferns; they simply produce sex organs and sperm and eggs rapidly, utilizing food in the spores. Many inhabitants of dry rock cliffs, especially in the maidenhair family, Adiantaceae, have developed a modified type of life cycle known as apogamy, in which fertilization is bypassed. This life cycle is also believed to foster quick reproduction in connection with brief damp periods; the gametophytes grow quickly, with buds developing directly into sporophytes; thus, free water is not required for swimming sperm.

Apogamous reproduction

Parasites and animals that feed upon ferns do not seem to be numerous, although the information available is not complete. Fungi infect ferns, some of them producing soruslike (*i.e.,* resembling the sorus, the sporangium cluster of ferns) dark bodies, or sclerotia. Snails and slugs commonly attack young, uncurling fronds (leaves) of some species, and various beetles have been observed to graze upon ferns. Partially eaten or insect-damaged fronds are not commonly observed in most fern species, however, which suggests that they may contain repulsive substances that ward off grazers. Many ferns have chemical compounds similar or identical to molting hormones of insects, and these may play a role in protecting the plants from major insect damage.

Although the sporophyte is long-lived, the fern gametophyte is usually ephemeral. It develops in a microenvironment characterized by little competition from other plants (including even mosses and algae); exposed humus, decomposing plant materials, or fresh mineral surfaces; deep to moderate shade; and a humid atmosphere. Even ferns whose sporophytes tolerate sun and drought tend to have these requirements for their gametophytes. On rocks, for example, the gametophytes form in protected crevices in which light is minimal and moisture maximal. Because of their requirements for exposed soil, development of fern gametophytes is promoted by damage to mature vegetation, such as fallen trees in the forest, flooding, and deep erosion. Prothallia are observed in nature most commonly upon shaded soil banks in forests and along streams and upon rotting logs.

As the bulk of reproduction of ferns is probably vegetative, taking place in the sporophytic stage, the presence of a large stand of a particular kind of fern results not so much from sexual reproduction by gametophytes as from clone formation by rhizomes and in some cases by root or leaf proliferations. In fact, vegetative reproduction probably accounts for the bulk of fern plants in the world; the sexual cycle, including spores and independent gametophytes, is probably important primarily in invading new

Drawing by M. Pahl

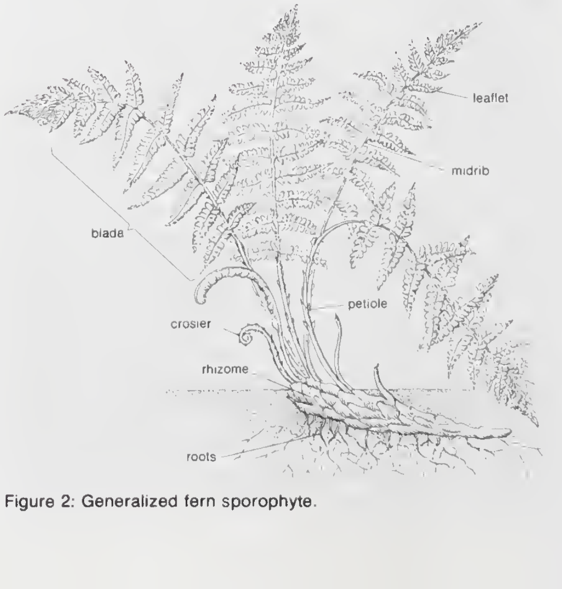


Figure 2: Generalized fern sporophyte.

habitats, extending the plant's geographic range, and creating ever-so-slight variations through rearrangement of the genetic material during meiotic cell division immediately preceding spore production.

FORM AND FUNCTION

Spore. The fern spore—a single living cell, usually protected by a thick wall—is the main source of population dispersal, being readily carried by wind. Ferns display a wide diversity of spore types in terms of shape, wall structure, and sexuality, and these types prove to have great value in determining taxonomic relationships. The full functional significance of the different types, except on the grossest scale, is not yet fully understood; for example, the minute differences in sculpturing of the outer wall surface do not, in the present state of knowledge, appear to have functional significance.

The tetrahedral spore

Shape. The basic spore shape among ferns is tetrahedral; the proximal face (the one facing inward during the tetrad, or four-cell, stage following reduction division, or meiosis) is made up of three sloping planes, and the distal, or outer, face consists of a single rounded surface. The tetrahedral structure is commonly obscured in so-called globose spores, the walls of which are thin and soft. Typically, the wall is composed of exospore (outer spore layer) only, there being no additional jacket, or perispore. The wall may be either unsculptured and smooth or provided with a variety of sculptured patterns. The tetrahedral spore is formed by simultaneous division of the products of the spore mother cell.

In contrast, the bilateral spore type of many fern species is formed by successive cell divisions of the spore mother cell. Where the tetrahedral spore possesses a triradiate scar on the proximal face—corresponding to its contact with three other spores in the tetrad—the bilateral spore has only a narrow, linear scar running parallel to the long axis. Most bilateral spores in ferns are bean-shaped and jacketed by a perisporial layer, a distinctive covering of the outer wall.

Size. Most ferns are homosporous, each plant having spores of one shape and size, usually 30 to 50 micrometres in length or diameter, although some reach more than 100 micrometres. A few fern families, however, have dimorphic spores, small ones (microspores) and large ones (megaspores). The gametophytes of ferns with dimorphic spores are endosporous; that is, they do not emerge in germination and fail to grow beyond the confines of the spore walls. Photosynthesis is essentially lacking, the food being stored in the spore. The microspores produce sperm in antheridia, and the megaspores produce eggs in archegonia. The vegetative phase of the gametophyte in these forms has been practically eliminated, and the developing embryo in the megaspore lives on stored food materials. The differentiation between male and female gametophytes ensures cross-fertilization. This set of conditions, known only in the families Marsileaceae, Azollaceae, and Salviniaceae, is called heterospory.

Wall. Spore walls may be thick or thin. Thick-walled spores are capable of surviving for a number of years, in some cases up to several decades. Sporocarps (masses of sporangia) of waterclover (*Marsilea*) 100 years old have been successfully germinated. Most natural germination of fern spores (except for water ferns) occurs on exposed damp surfaces of rock, soil, or dead plant materials.

A number of fern genera (*e.g.*, *Osmunda, Grammitis, Hymenophyllum, Trichomanes, Matteuccia*) possess thin-walled spores. In practically all known examples, such thin-walled spores are also green-pigmented, being provided with chloroplasts. Such spores are common among rain forest genera; they are often short-lived and require a short time for germination. Spores of *Hymenophyllum, Trichomanes,* and *Grammitis* remain viable only a few days, those of *Osmunda* a few months.

Growth of the gametophyte

Gametophyte. When the spore wall cracks under appropriate moist conditions, the fern gametophyte is formed. Emerging from the spore at the time of germination are a nongreen rhizoid (rootlike organ), which attaches the plant to the growing surface, and a green single cell—the mother cell that gives rise to the rest of the gametophyte. At first, in most homosporous ferns, growth is in the form of a single filament, and it may continue in this fashion if lighting conditions are weak. If lighting is optimal, however, the gametophyte becomes a two-dimensional sheet of cells and later a layered, three-dimensional structure. The apical cell, which initiates growth, is soon replaced by a growth zone, or meristem, which, as a result of the directions of cell division and enlargement, comes to lie in an apical notch in the gametophyte, surrounded on either side by the prothallial wings—flat, platelike protrusions, one cell thick. The average size of the gametophyte at the time of fertilization is approximately 2 to 8 millimetres (0.08 to 0.32 inch) long and up to 8 millimetres wide.

Specialized forms. From a basic type of gametophyte, somewhat like that just described, a number of highly specialized forms have evolved that are characteristic of certain genera. Ribbonlike gametophytes are known especially among tropical rain forest ferns such as *Vittaria, Grammitis,* and *Hymenophyllum* and are usually irregularly and extensively branched, forming large masses of intertwining ribbons. Some of these are actually more abundant than their corresponding sporophytes in certain localities (*e.g.*, the Appalachian Mountains, where "pure cultures" of gametophytes totally lacking sporophytes have commonly been found). Filamentous (threadlike) gametophytes are known in the genera *Trichomanes* (Hymenophyllaceae) and *Schizaea* (Schizaeaceae).

Tuberlike gametophytes occur in several groups—*e.g.*, the families Ophioglossaceae (all members), Schizaeaceae (*Actinostachys*), and Stromatopteridaceae (*Stromatopteris*). All are nongreen underground plants that have close associations with fungi and are therefore assumed to be saprophytic (*i.e.*, dependent for nutrition upon rotting organic material in the soil). They commonly occur 5 to 10 centimetres (2 to 4 inches) deep in the ground.

In heterosporous ferns the endosporous gametophytes are

Drawing by M Pahl

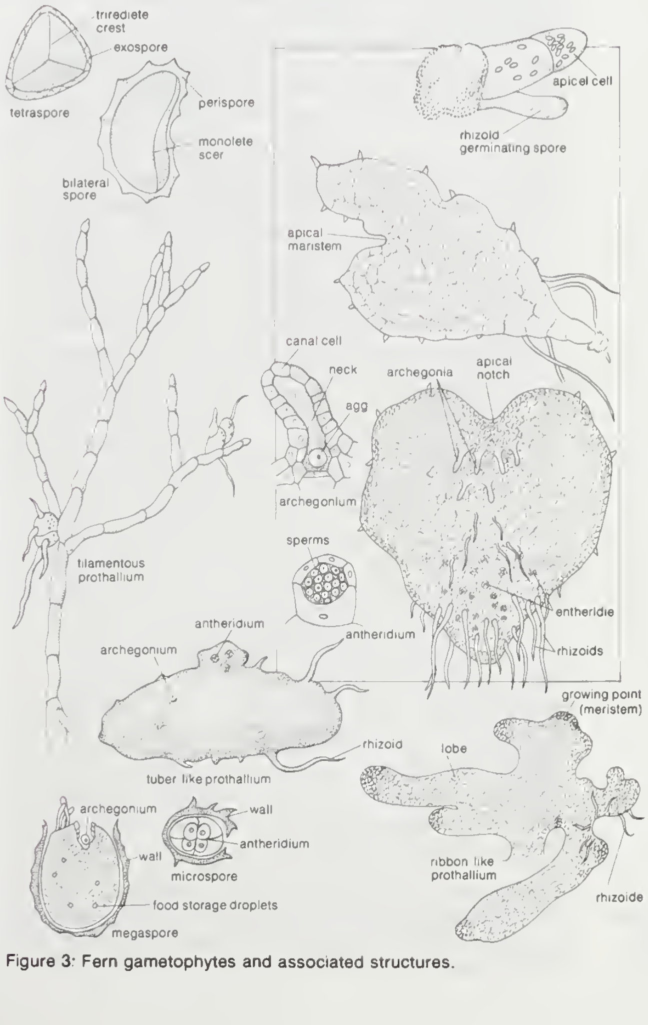


Figure 3: Fern gametophytes and associated structures.

much reduced. The male gametophyte in the microspore is made up of the equivalent of one antheridium and its complement of sperm. The female gametophyte, although considerably larger in size, is equally reduced in a morphological sense, the greater space within the megaspore being filled by stored nutritive materials and tissues formed around the base of the female sex organs.

Vegetative reproduction. Vegetative propagation of some photosynthetic fern gametophytes is accomplished by continued growth and fragmentation, but this does not spread the gametophyte very far. Some ferns (*Vittaria, Grammitis,* and the family Hymenophyllaceae) produce specialized filaments, or gemmae, that break off and are carried away by water droplets, wind, or possibly crickets to initiate new colonies.

Structure of the archegonium

Sexual reproduction. The sex organs of ferns are of two types. The sperm-producing organ, the antheridium, consists of a jacket of sterile cells with sperm-producing cells inside. Antheridia may be sunken (as in the families Ophioglossaceae and Marattiaceae) or protruding. They vary in size from those with hundreds of sperm to those with only 12 or so. The egg-producing organ, the archegonium, contains one gamete (sex cell), which is always located in the lower, more or less dilated portion of the archegonium, the venter. The upper part of the archegonium, the neck, consists of four rows of cells containing central neck cells. The uppermost of the neck cells are the neck canal cells; the lowest cell is the ventral canal cell, which is situated just above the egg.

Fertilization is attained by the ejection of sperm from antheridia. The sperm swim through free water toward simple organic acids released at the opening of the archegonium, the neck of which spreads apart at the apex, permitting the neck cells to be extruded and the sperm to swim in and penetrate the egg. The sperm are made up almost entirely of nuclear material, but their surface is provided with spiral bands of cilia—hairlike organs that effect locomotion. When the egg is fertilized, the base of the neck closes, and the embryo develops within the expanding venter.

Embryo. Within the archegonial venter the zygote undergoes characteristic cell divisions to form the embryo, which remains encapsulated in the gametophyte until it breaks out and becomes an independent plant. The pattern of development in most ferns is a distinctive one, and indeed only in the classes Ophioglossopsida and Marattiopsida—in the *Botrychium* subgenus *Sceptridium* and in all species of the family Marattiaceae thus far studied—are found conditions of embryonic development resembling those of seed plants. Here the first division of the zygote is transverse. The inner cell grows inward, producing the stem and first leaf, and the outer cell divides to form a foot, a mass of tissue that exists as part of the embryo and disappears when its function, presumably absorption, is completed. The root appears later within the stem and grows outward. In all other known ferns the zygote divides neatly into four quadrants, the first division approximately parallel to the long axis of the archegonium and the following division at right angles. This results in initial cells that give rise to four organs: the outer forward cell (*i.e.,* toward the growing apex of the gametophyte and the neck of the archegonium) becomes the first leaf, the inner forward cell the stem apex, the outer back cell the first root, and the inner back cell the foot. Thus, the majority of ferns tend to have a precise arrangement of their organs and the divisions that produce them in the embryo.

The young sporophytes of ferns remain attached to the gametophytes for varying lengths of time, absorbing nutrients from the gametophyte through the foot. Once the sporophyte has developed independent existence and the root has penetrated the soil, the gametophyte soon shrivels.

Ant chambers in fern stems

Stem. Fern stems vary from the tall, narrow trunks of certain tree ferns that reach 25 metres (80 feet) tall down to creeping rootstocks, or rhizomes. Rhizomes are the most common stem form. The majority of them grow horizontally upon or just beneath the surface of the soil. Some stems are so narrow as to be threadlike, as in many tropical epiphytic ferns. A few ferns in different parts of the world have evolved radically specialized stems containing chambers in which ants take up residence; the role of the ants in the lives of these ferns is unknown, but it may be for protection against other insects. Vinelike ferns are common, but shrubby ferns are extremely rare.

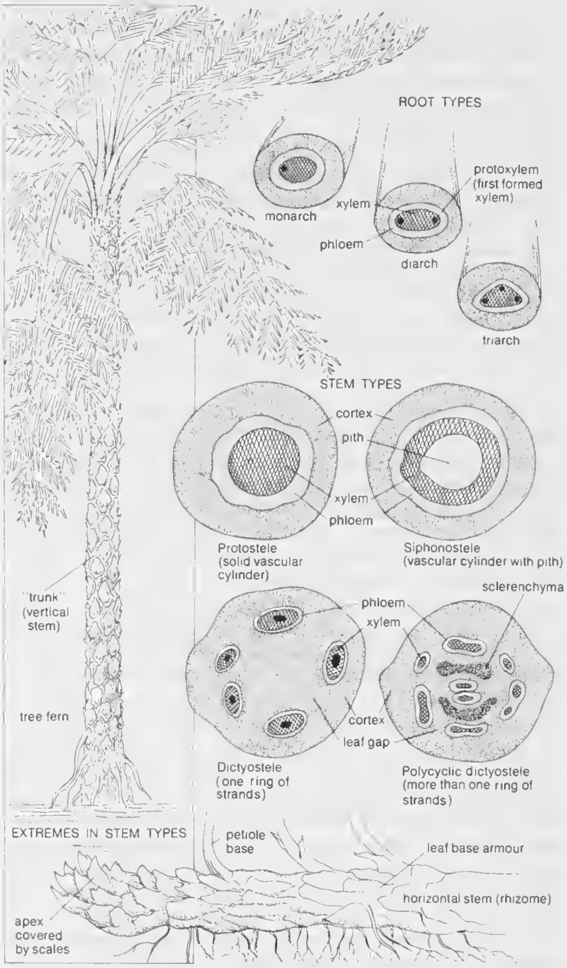

ROOT TYPES
protoxylem (first formed xylem)
xylem
monarch
phloem
diarch
triarch
STEM TYPES
cortex
pith
xylem
phloem
Protostele (solid vascular cylinder)
Siphonostele (vascular cylinder with pith)
sclerenchyma
phloem
xylem
"trunk" (vertical stem)
tree fern
cortex
leaf gap
Dictyostele (one ring of strands)
Polycyclic dictyostele (more than one ring of strands)
EXTREMES IN STEM TYPES
petiole base
leaf base armour
horizontal stem (rhizome)
apex covered by scales

Figure 4: Stem structures in ferns.
Drawing by M. Pahl

Stem growth is initiated by one to several large apical cells. These are usually well protected by various types of hairs or scales and by the overarching embryonic leaves. Leaves and leaf bases play a major role in the protection of fern stems, and many stems are said to have a leaf armour. Such stems are densely covered with old, sclerified leaf bases, which increase the apparent size of the stem many times. The old leaf bases may serve as protection or as food storage organs. In most species the stems are indeterminate in growth and thus can theoretically continue to grow indefinitely. Annuals—short-lived species that complete development, shed spores, and die in a single growing season—are exceptional; only one or two examples are known.

Surface structure. Whether covered with leaf armour or not, the surface of the fern stem is protected by an epidermis, or "skin," a single layer of epidermal cells, which are more or less flat cells with thick outer walls. Most fern stems also are covered with a protective indument, consisting of hairs or scales; these are so distinctive that they are valuable in identification and classification. The indument includes such diverse types of epidermal emergences as simple glands (unbranched, one- to several-celled hairs with a headlike cluster of secretory terminal cells); simple (unbranched), nonglandular hairs; dendroid hairs (branching filaments); and scales (flat cell plates) of many patterns. Scales (also known as paleae) are defined as a cell plate two or more cell rows wide, at least at the base, whereas hairs generally have a single row of cells. Transitional states are also known.

Cortex. The cortex is the region outside the vascular cylinder but below the surface of the stem. It is composed mostly of storage parenchyma cells (a relatively generalized cell type). Rooting animals, such as pigs, occasionally dig up fern rhizomes for the starchy materials contained in them. There is a strong tendency for the outermost cortical cells to become darkly pigmented and thick-walled.

Types of steles

Vascular tissues. The steles—cylinders of vascular tissues in the centres of fern stems—exhibit somewhat diverse patterns. Most common ferns possess a "dictyostele," consisting of vascular strands interconnected in such a manner that, in any given cross section of stem, several distinct bundles can be observed. These are separated by regions filled with parenchyma cells known as leaf gaps. There are, however, numerous "siphonostelic" ferns, in which the gaps do not overlap and a given section shows only one gap, and some "protostelic" ferns, in which no gaps at all are formed. Complex stelar patterns are known in some species, as in the common bracken fern (*Pteridium*), which has a polycyclic dictyostele, one in which one stele occurs within another stele; large strands of fibrelike cells running between them form mechanically specialized hard tissue, or sclerenchyma.

Root. Fern roots are generally thin and wiry, although some are fleshy and either slender (in the Ophioglossaceae) or as much as 13 millimetres (0.5 inch) in diameter (*e.g., Acrostichum, Marattia*). The relation of the roots to the stems is a valuable identification tool. For example, in certain tree ferns (*e.g., Cyathea, Cibotium*) and in the royal ferns (*Osmunda*) the entire stem surface is covered by masses of roots. If large enough, the dead tangles of tree fern roots can be cut with a saw into various shapes suitable for attaching epiphytic greenhouse plants, and pieces of such root masses have proved to be useful in horticulture for cultivating orchids and bromeliads. Because of the massive destruction of tree ferns for this purpose, the importation of tree fern logs into many countries is now prohibited. Certain tropical ferns have elaborately hairy roots whose surfaces are covered with locks of silky golden or brown root hairs.

Leaf. *Shapes.* The leaf plan in practically all ferns is pinnate—that is, featherlike with a central axis and smaller side branches—and this is considered to be the primitive condition because of its widespread occurrence. From this basic type there has evolved a broad diversity of forms. Some ferns have palmate leaves (with veins or leaflets radiating from one point), and some, such as the staghorn ferns, have secondarily evolved falsely dichotomous leaves. In some genera (*Lygodium, Salpichlaena*) the main leaf axis (rachis) twines about on shrubs and small trees, sometimes reaching 20 metres (65 feet) in length.

Whether a given leaf is divided into segments (compound) or is undivided (simple) is of considerable value in identification of similar fern species. The difference between divided and undivided leaves is not a profound one, however, and closely related species commonly differ from one another in this respect.

Variations in leaf divisions

The extent of division in fern leaves, or fronds, ranges from that in which the leaf margins are merely so deeply lobed as to have narrow-based segments to that of having obviously stalked leaflets, or pinnae. The pinnae themselves may also be lobed or truly divided with stalked segments as well; and the resulting segments, the pinnules, may also be lobed or divided. Depending on the degree of cutting, fronds are described as simple, once divided, twice divided, thrice divided, and so on. Some ferns are known in which the fronds are five times compound, making them exceedingly delicate, with segments so small as to be almost hairlike.

Venation. Generally, the patterns of the leaf veins, or vascular bundles (which can be seen readily by holding the specimen up to a strong light), are pinnate and the veins are free; that is, they all diverge and never coalesce, either along their sides or at the ends. Nevertheless, there are numerous fern groups in which netted, or reticulate, venation is found. These have vein patterns like those of other ferns for the most part, except that various systems of networks and areolae (areas enclosed within loops of veins) have developed between the major, pinnately arranged veins. There are many reticulate patterns known. One of the more striking is that in which each loop or areola contains one or more free included veinlets, as seen in various members of the family Polypodiaceae. Another is the herringbone pattern, believed to result from an evolutionary concrescence (growing together) of pinnae, as shown by certain tree ferns (*Cyathea*), lady ferns (*Athyrium*), and marsh ferns (*Thelypteris*).

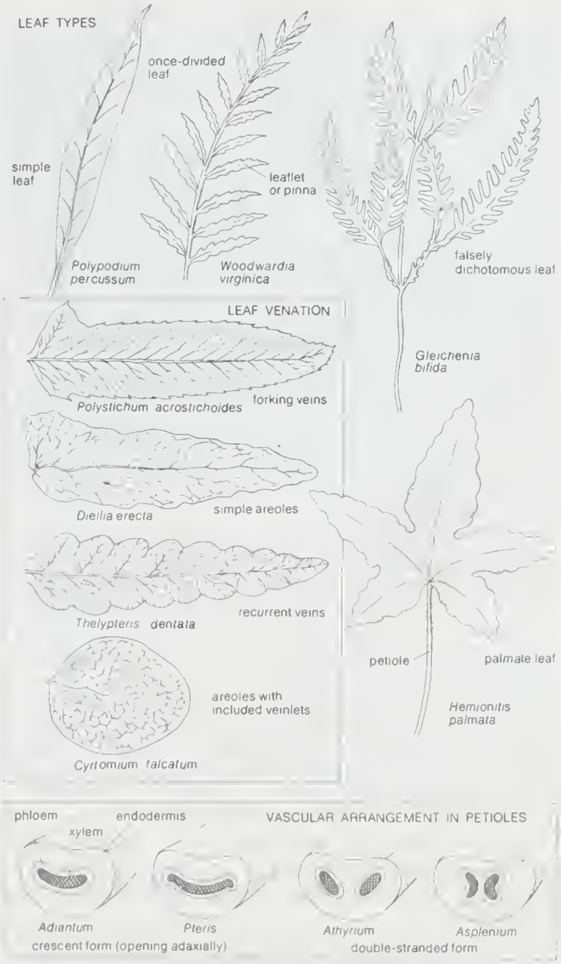


Figure 5: Fern leaves, showing leaf types, leaf venation, and internal petiole vascularization.

Drawing by M. Pahl

Leaf position

Leaf stalk. Fern leaves vary in the relationship of the petiole, or leaf stalk, to the blade (the expanded part of the leaf). Many strap-shaped leaves essentially have no petiole and are described as sessile; broad, ovate, or triangular leaves commonly have a pronounced leaf stalk, called a stipe, and are termed petiolate or stipitate. Narrowly elongated leaves in ferns are usually erect, spreading, or, in certain epiphytes, pendent. Leaves that are broadly ovate or triangular tend to be borne at right angles to the incident light. Broad-leaved ferns thus become more or less bent at the blade base, with an arch at the top of the petiole.

Anatomically, the petioles or stipes of fern leaves show nearly as much diversity in cross-sectional pattern as do the stems. The simplest vascular strands of fern petioles are commonly crescent-shaped, single bundles. In more and more elaborate petiolar patterns, the crescent takes on the form of the Greek letter omega (Ω), opening adaxially (*i.e.,* upward or toward the central axis of the plant). The latter shape, with many variations, occurs widely among ferns, especially those considered on other grounds to be primitive. Double-stranded ferns (the omega now divided into two parts and unconnected below) are usually associated with more specialized genera (*e.g., Athyrium, Thelypteris*). Any of the generalized patterns may exist as broken-up strands; the separation is commonly associated with size, small leaves having only three to nine strands, large leaves of tree ferns having many. Petiolar vascular bundle shapes have been found to be so definitive as to have a certain value in separating fern genera and families.

Tissues. At the tissue level the leaf blades have certain differences from those of other plants, but the same general picture prevails. There are an upper and a lower epidermis, the latter with many stomates (microscopic pores). In between, the mesophyll is usually composed of cells with large intercellular spaces. In thicker leaves, the upper mesophyll is composed of palisade cells—elongated cells arranged parallel and oriented with the long axis vertical to the leaf surface. The veins range from the massive major midribs, or rachises, which have well-defined xylem, phloem, pericycle, and endodermis, to the delicate capillaries represented by little more than a single file of tracheids that sometimes ends in a cluster of somewhat modified tracheids. Underneath a sporangium or sorus, the veins may become dilated and multilayered (so-called fertile veins). In many ferns all or nearly all of the photosynthesis is accomplished by the epidermis, the mesophyll having been eliminated in evolution. An example is the common maidenhair fern (*Adiantum pedatum*), the blade of which, between veins, is mainly made up of only two layers, the upper and the lower epidermis, in which most photosynthesis occurs.

Indument on fern leaves

The indument of fern leaves may be like that of the stem, but usually the hairs or scales are fewer, more widely separated, and smaller, or they may be of a different type. Numerous ferns are described as having glabrous (bald) leaf blades, but many of these actually have at least a few microscopic hairs, which are usually glandular and appressed to the blade surface.

Comparisons with leaves of other plant groups. The fern leaf, or pteridophyll, differs from the "true leaf" (euphyll) of the flowering plants in its vernation, or manner of expanding from the bud. In the ferns, vernation is circinate; that is, the leaf unrolls from the tip, with the appearance of a fiddlehead, rather than expanding from a folded condition. It also differs in its venation, which usually is free or simply reticulate rather than being highly complex and made up of areolae containing numerous branched, free-ending veinlets.

Fern leaves differ from the leaves (sphenophylls) of conifers and horsetails in that fern leaves usually display a well-developed central midrib with lateral vein branches rather than a dichotomous, midribless pattern or a simple vein in a narrow, needlelike, or straplike leaf. Although a few ferns that have narrow leaves also have only a single central vascular strand (*e.g.*, certain species of *Schizaea*), these can usually be distinguished readily from the scale-like or awllike leaves (microphylls) of club mosses on the basis of other characteristics, such as the position of the sporangia and the mode of leaf development. A few genera of ferns (*e.g.*, sword ferns, *Nephrolepis; Jamesonia; Salpichlaena;* and climbing ferns, *Lygodium*) have members with more or less indeterminate (*i.e.*, continuous) leaf growth accomplished by periodically quiescent buds. Fern leaves, however, are mostly determinate; that is, they stop growing when they reach maturity. Leaves grow from apical cells in most ferns, and these delicate embryonic cells are protected by the curled-over spiral of the crosier (unrolling leaf tip) and by hairs or scales. When the blade formation is complete, there is no longer an embryonic tip.

In overall length, mature leaves vary from 1 or 2 millimetres (0.04 or 0.08 inch) in certain filmy ferns (Hymenophyllaceae) to 30 or more metres (100 feet; family Gleicheniaceae). In terms of overall size, the most massive frond is that of the elephant fern (*Angiopteris*), with fronds more than 5 metres long and petioles 15 centimetres in diameter.

Drawing by M Pahl

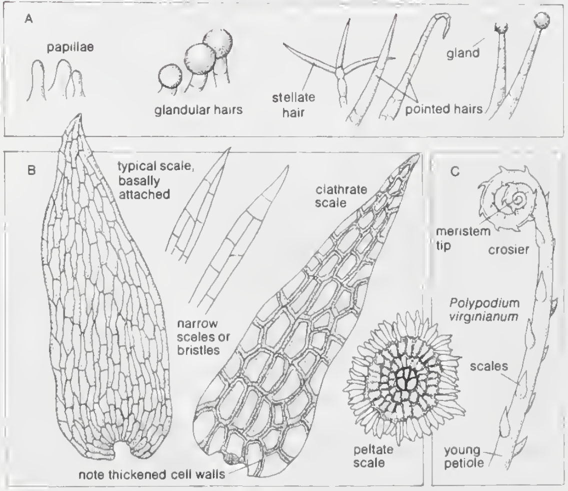


Figure 6: *Representative surface structures of fern leaves.* (A) Types of hairs. (B) Scale types. (C) Uncurling leaf, or crosier, showing circinate vernation and surface scales.

Types of spore-producing structures

Sporangium and sorus. *The sporangium.* The spore cases, or spore-producing structures, in ferns range from globose, sessile (nonstalked) organs more than 1 millimetre in diameter down to microscopic stalked structures, the capsules of which are only 0.3 millimetre in diameter. The former are known as eusporangia, the latter as leptosporangia. Eusporangia occur in the classes Ophioglossopsida and Marattiopsida, and leptosporangia occur in the majority of the species in the class Filicopsida. There are, however, many intermediate forms between the two types of sporangia, and these are known in various primitive species of the Filicopsida, such as members of the family Osmundaceae.

The capsule wall in eusporangia tends to be relatively massive, made up of two layers or more. In leptosporangia, on the other hand, the wall is thin and, at least at maturity, composed of one layer of cells. Opening of the capsule in eusporangia, such as those of the genus *Botrychium*, is accomplished by separation along a well-differentiated line of dehiscence (opening); but in most typical leptosporangia, except for a few stomial ("mouth") cells that separate along one side, the process of dehiscence tears the cells apart more or less irregularly.

The opening process in eusporangia is the result of a generalized stress on drying walls, the cells of which are differentially thickened. There is no mechanism to throw the spores, and they are simply carried away by the wind. In contrast, leptosporangia display more or less specialized bows, or annuli, usually consisting of a single row of differentially thickened cells. Apparently, the mechanical force for opening and for throwing the spores derives entirely from these annular cells; all the other capsule cells are thin-walled and unmodified. The stresses imposed by the drying of the annular cells result in the collapse of the outer sides of the cells, thus straightening out the annulus and ripping the soft lateral cells of the capsule apart. As the annular cells continue to be deformed by the cohesive forces of the increasingly tense water molecules within, the spore case completely opens. Finally, the cohesive capacity of the water molecules is exceeded, the water film between the outer walls of the annular cells breaks, and the entire annulus snaps back to its original position, tossing the spores into the air.

Most primitive sporangial types are stalkless, or sessile. If a stalk is present at all, it is merely a slightly raised multicellular area at the base of the sporangial capsule. In typical leptosporangia, however, there commonly are well-developed stalks, and these are often extremely long and narrow (*e.g.*, as in *Davallia* and *Loxoscaphe*), made up of only one or two rows of cells and often 1.7 to 2 millimetres in length.

The trend in sporangial evolution is evidently from solitary large capsules to more and more elaborate groupings of smaller sporangia. These changes are accompanied by the appearance of such refinements as paraphyses and indusia. Paraphyses are sterile structures that grow among or on the sporangia. Indusia are papery, tentlike structures that cover the sori (clusters of sporangia).

The sorus. Hand in hand with the reduction in size of single sporangia are seen more and more complex aggregations of sporangia known as sori. The meristematic area—the region of new cell growth—that produces them may continue its activities over a number of weeks, producing sporangia of all ages, older ones being pushed aside as new ones mature in their turn. When sori develop on the leaves of house ferns, they are often mistaken for tiny insects (young stages) or a fungus disease (older stages) rather than recognized as organs necessary for the normal reproduction of the plant.

Evolution of the sorus

The stages in progressive evolution of sori can be depicted as follows: (1) simple clusters of sporangia, these more or less coalesced (family Marattiaceae) or separate (Gleicheniaceae), all of them maturing at the same time, (2) gradate clusters of sporangia, the outermost ones maturing first, the innermost last, and (3) mixed clusters of sporangia, all ages present, the younger ones arising from the same meristematic zones as the older ones. The adaptive significance of this change is probably related to the duration of spore production, the mixed character of the more advanced sori extending the period beyond that of solitary sporangia or of simple, simultaneously maturing sori.

Drawing by M. Pahl from (top) F. C. Steward, *About Plants: Topics in Plant Biology* (1966), Addison-Wesley, Reading, Mass., and (bottom right) P. Martens & N. Pirard, "The Glandular Organs of *Polypodium virginianum*," *La Cellule*, 49(3) 383–406 (1943)

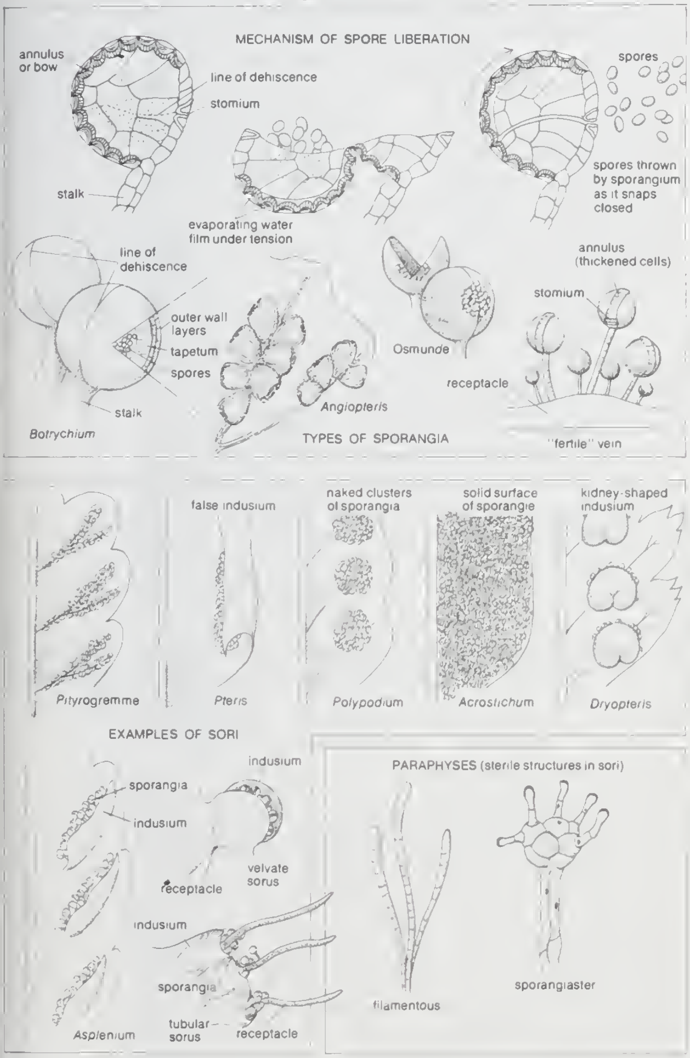


Figure 7: Fern sporangia, their arrangement into sori, and other structures found in the sorus.

Sporangia and especially sori have traditionally provided the most important characters for fern classification. Indeed, many unrelated ferns were once classified together because of what are now believed to have been coincidental convergences in soral structure. Between one-half and two-thirds of the species of ferns have one or another of the following six soral arrangements: (1) A linear arrangement of sporangia along veins, avoiding the leaf area between the veins, is found in many fern genera, especially in the genus *Pityrogramma*. (2) A line of sporangia along the leaf edge, protected usually by a rolled-over and modified laminar margin, is represented by *Pteris*. (3) Round and naked sori (*i.e.,* without an indusium) are found in *Polypodium*. (4) An arrangement of large sori that usually expand over the entire undersurface of the blade or pinna is represented by *Acrostichum*. Such sori probably arose by the fusion of smaller clusters of sori. Of the many arrangements of indusiate sori (*i.e.,* sori that are protected by indusia, or special scalelike structures), two of the most widespread are (5) a linear or oblong sorus along a vein covered from one side by a narrow indusium, which is represented by *Asplenium*, and (6) a sorus that is round but covered with a kidney-shaped or shieldlike indusium, which is represented by *Dryopteris*.

Variety in structure of indusia

The indusium. Protection of the sporangial cluster from exposure, drying, and other hazards is accomplished in various ways, such as by the formation of the sori in grooves or pockets or by the production of various forms of covers. One is the so-called false indusium, a rolled-over leaf margin under which sporangia form and mature. The true indusium is a separate and unique formation, the structural origins of which are not clear, that constitutes a more or less papery covering over the sorus. A widespread type of indusium among members of the family Cyatheaceae is one shaped like a cup, which arises around the base of the sorus, often enclosing the sorus until the sporangia are mature (*e.g., Cyathea*). In some genera, marginal sori are protected by a two-lipped, or valvate, indusium (*e.g., Dennstaedtia, Dicksonia, Hymenophyllum*). When sori fuse laterally to form continuous lines, or coenosori, any indusia also tend to fuse.

Paraphyses. Approximately one-third of fern species have paraphyses of one type or another. These are sterile hairs or scales intermixed with the sporangia, and they are, like indusia, believed to perform a protective function. Paraphyses usually are hairs or modifications of hairs that arise among the sporangia or on the sporangial stalk or capsule. In various genera of ferns, the paraphyses have proved to be helpful sources of taxonomic data.

CYTOGENETICS

Chromosome numbers and polyploidy. The study of chromosomes, hybrids, and breeding systems has revealed much of value in understanding ferns. The chromosomes of ferns tend to have high base, or x, numbers, ranging from approximately 20 to 70, with the majority between 25 and 45. The familiar genus *Osmunda*, for example, has $x = 22$, *Pteris* has 29, *Asplenium* 36, *Dryopteris* 41, *Botrychium* 45, and *Pteridium* 52. Among homosporous ferns, exceptions to the rule of high chromosome numbers are rare; in one species of filmy fern (*Hymenophyllum peltatum*), $x = 11$, the lowest number reported. Among heterosporous ferns, however, the situation is conspicuously different, and all have low base numbers (*Marsilea*, $x = 10$, 13, or 19; *Salvinia*, $x = 9$; *Azolla*, $x = 22$).

The explanation for the difference traditionally adopted by cytologists is that the high numbers in homosporous ferns arose from polyploidy, the repeated duplication of whole sets of chromosomes. Indeed, some workers regard homosporous ferns as nearly 100 percent polyploid. An alternative hypothesis should also be considered, however; namely, homosporous ferns were primitively high numbered, and heterosporous ferns derived their low numbers through reduction. Apparently the ferns do not need all their chromosomes; recent evidence has shown a considerable degree of "gene silencing."

The base chromosome numbers (indicated by the symbol x) have been used for classification purposes. Commonly, the base number is uniform for a genus or family, or it ranges around a given number. More rarely, the number varies drastically, as in the genus *Thelypteris*, which has x numbers ranging from 27 to 36, or *Lindsaea*, with x numbers from 34 to about 50. So much variation in the chromosome base number suggests that the "genus" concerned may be unnatural or that it may be very ancient, with intermediate numbers having disappeared (*e.g., Dennstaedtia*), or that it is in a state of active evolution (*Thelypteris*).

Simple polyploid series—multiples of the base number—are prevalent among ferns, and a few species are reported to have forms or races that are diploid (with two times the base number of chromosomes), tetraploid (four times), and hexaploid (six times). For example, the fragile fern, *Cystopteris fragilis*, has races with the number of chromosomes per nucleus in the sporophyte generation—represented by $2n$—equal to two, four, and six times the base number of $x = 42$; or $2n = 84$, 168, and 252. Species with both diploid and tetraploid forms are common, especially among widespread, abundant ferns. In most cases the cytological races are differentiated on quantitative characters, especially the sizes of such cells as spores, epidermal guard cells (cells next to stomates), and hair cells.

Hybridization. In certain temperate fern genera, such as spleenworts (*Asplenium*), wood ferns (*Dryopteris*), and holly ferns (*Polystichum*), hybridization between species (interspecific crossing) may be so frequent as to cause serious taxonomic problems. Hybridization between genera is rare but has been reported between closely related groups. Fern hybrids are conspicuously intermediate in characteristics between their parents, and simple dominance of single characters is unusual. Occasionally, when the interspecific crosses involve strongly different characteristics, the hybrid displays an irregularity in expression of these characteristics, often involving marked asymmetry. The majority of hybrids are sterile and reproduce, if at all, only by vegetative propagation.

Reproduction in sterile fern hybrids

Reproduction in sterile fern hybrids is also accomplished by the process of apogamy, in which spores possessing the same chromosome complement as the sporophyte are produced (normal spores have only half the chromosome number as the parent plant cells). These unreduced spores (with the $2n$ number of chromosomes) are viable and germinate into normal-appearing gametophytes that may or may not form sex organs. The hybrid gametophytes do not, however, undergo normal sexual fusion. Instead, the meristematic (cell-producing) region of the prothallium simply buds off a new sporophyte, and there is a direct conversion from gametophyte to sporophyte generation.

In most fern hybrids the spore mother cells are unable to form bivalents (chromosome pairs) at meiosis, and reduction division results in irregular, deformed, and inviable spores. In the sporangia of most apogamous ferns, however, automatic doubling of chromosomes occurs by endomitosis (duplication of chromosomes without formation of two nuclei), and each of the spore mother cells has a restitution nucleus—one with doubled chromosomes. In these doubled sporangia there are, therefore, only 8 spore mother cells rather than the usual 16, and they undergo meiosis, producing viable diploid spores. Apogamous ferns are known in a number of genera of higher ferns in various families, including *Adiantum, Asplenium, Cheilanthes, Dryopteris, Pellaea, Polypodium,* and *Pteris.*

Allopolyploid hybrids

Besides apogamous hybrids, there are numerous demonstrated or suspected "allopolyploid hybrids," which are believed to have originated by doubling of the chromosomes of sterile crosses. These are intermediate in their characteristics between well-known parental species and behave like normal, divergent species, alternating sporophytes with gametophytes and undergoing normal meiosis and fertilization. Genera with frequent hybridization often exhibit a variety of chromosome numbers that are multiples of the generic base number. One of the best examples is the tropical genus *Anemia,* with the base number of 38 and species with 76, 114, 152, 190, and 266.

Both apogamous and allopolyploid hybrids may enjoy wide geographic ranges and occur in as great abundance as normal species. Both types of hybrids are also capable of creating additional hybrids by backcrossing (to the parent species) or by crossing with other species. In apogamous ferns it is assumed that the sperm are generally viable and capable of fusing with eggs of other, normal species. In total, hybrids—sterile, apogamous, and allopolyploid—may make up as many as 25 percent of the different kinds of ferns in a given flora.

Curiously, in spite of the high number of ferns that are epiphytic (growing on trees), nearly all the fern hybrids are terrestrial or epipetric (growing on rocks); hybridization is very rare among epiphytes. The reason for this phenomenon is not yet clear; it could be simply that the mosses and decaying leaves on tree trunks and branches may keep the individual gametophytes apart, whereas on muddy banks gametophytes of different species may be in close proximity.

ORIGIN AND EVOLUTION

Fossil record. Fernlike characteristics are known to be combined in numerous fossils coming from geologic strata as old as the Devonian (beginning 408 million years ago). The Carboniferous Period (360 to 286 million years ago) was a time of great evolutionary experimentation in ferns, but nearly all those groups are now extinct. Modern ferns, however, are relatively uniform in basic structure, and they share a large number of characteristics, combined in a distinctive way. All the living families, with the possible exception of the Ophioglossopsida and the Marattiopsida, possess a ground plan of correlated characteristics that seems clearly to bind them together as an assemblage that is monophyletic (*i.e.,* having one evolutionary line). In spite of this, the ferns still display wide variation.

The norm of modern ferns is so distinctive that the vast majority of them can be recognized immediately as members of this plant division. Nevertheless, various workers in the past, especially among paleobotanists, have singled out fossil fragments and speculated that they represent fern ancestors, sometimes giving them such names as *Archaeopteris* (primitive fern) or *Protopteridium* (first fern). One or two extant genera can be traced to direct ancestors in the Carboniferous, but for the most part the fossil record shows no immediate ancestors of modern ferns; the first relatives of today's ferns in the fossil record are usually classifiable into living groups.

The earliest true ferns arose during Carboniferous times, or perhaps a few in Devonian, and have been classified in four families—Marattiaceae, Osmundaceae, Gleicheniaceae, and Schizaeaceae. Several extinct groups of the Carboniferous Period and the Permian Period that followed—Coenopteridaceae, Anachoropteridaceae, Tedeliaceae, Sermayaceae, Tempskyaceae—represent related lines of evolution, but there are no intermediate examples to show close ties with any of the modern families of ferns. The immediate ancestors of the extinct seed ferns (pteridosperms) may also have been the immediate ancestors of modern ferns, judging from numerous data on sporangial arrangements and shapes as well as on leaf anatomy. What used to be considered impressions of fern leaves from fossils dating from the Carboniferous have been shown in many cases to have borne seeds or to have been associated with seed-bearing plants.

Antiquity of modern ferns

By the time of the Mesozoic Era (beginning 245 million years ago), the modern fern families were well established, and there are fossil records of the families Osmundaceae, Marattiaceae, Schizaeaceae, Matoniaceae, Dipteridaceae, Adiantaceae, Cyatheaceae, Aspleniaceae, Marsileaceae, Azollaceae, and Salviniaceae.

Evolutionary development. Despite a relatively large number of theories, the actual origins of the vegetative organs of ferns are still unknown. It is usually suggested that the original fern stem was protostelic (its stele having no pith or leaf gaps), but this is not necessarily true of the immediate ancestor of modern ferns. In fact, it is conceivable that "eustelar" stems, with secondary growth (*i.e.,* growth in thickness, as in the stems of modern conifers and woody flowering plants), gave rise to modern fern stems through reduction and disappearance of the secondary growth and replacement of the stele by overlapping leaf traces (the vascular bundles from stele to leaf).

The leaf is equally or even more problematic as to its ultimate origin. Various hypotheses have been offered, of which the telome theory (that the leaf arose from fusions and rearrangements of branching stem systems) and the enation theory (that the leaf arose from simple enations, or outgrowths) are the two most popular. The true story seems to be lost in antiquity and perhaps will never be known. Modern fern leaves with their characteristic fiddleheads, acropetal growth (*i.e.,* "seeking the apex," the leaf tissues maturing from the base toward the tip, where the youngest tissues are produced), and pinnate structure, are nevertheless quite distinctive. They differ in numerous respects from sphenophylls, such as those of conifers and sphenophytes, and from euphylls, such as those of flowering plants. It is possible that these leaf types did not originate in the same way and even that different examples of each had different origins.

CLASSIFICATION

Annotated classification. The classification presented here is derived from J.A. Crabbe, A.C. Jermy, and J.T. Mickel (1975) and T.N. Taylor (1981). Numbers given for the species are only rough approximations of living groups.

DIVISION FILICOPHYTA (true ferns)

Spore-dispersed vascular plants with free-living sporophyte and gametophyte generations, the sporophyte dominant; leaves frondose, circinate (expanding by unrolling), possessing petiole (stalk), midrib, and blade, the latter or its subdivisions with free or netted venation patterns; sporangia solitary or variously soriate (in clusters) and borne upon the leaf blades or their modified parts, marginally or abaxially (on the undersurface, or surface facing away from the central axis of the plant); stems upright or creeping, mostly lacking secondary thickening, and bearing roots; meristems (growing points) protected by hairs or scales; gametophytes monomorphic (all with similar structure), exosporic (developing outside the spores), hermaphroditic (bisexual), and either surficial (growing on the surface) and photosynthetic or subterranean, saprophytic, and symbiotic with fungi; or gametophytes dimorphic (with two forms), endosporic, and antheridial (male) or archegonial (female).

Class Ophioglossopsida

Leaf divided into sterile and fertile segments, leaf base more or less clasping the stem; eusporangiate (with unstalked, globose sporangia); homosporous; plants mostly small and fleshy, occurring in early ecological succession; not known as fossils.

Family Ophioglossaceae (adder's-tongue, grape ferns, moonworts). Leaf blade undivided to forked or pinnately lobed or divided; veins reticulate (netlike) or free; sporangia borne upon unbranched "spike" or many-branched fertile structure arising from base of the blade or blade stalk; 3 genera (*Botrychium, Helminthostachys, Ophioglossum*) with about 55 species.

Class Marattiopsida (giant ferns)

Leaves pinnately divided, pulvinate (enlarged or swollen at attachment point of leaflets) in living genera, and with well-developed, fleshy stipules (appendages at leaf base); sporangia eusporangiate, in sori, or more or less coalescent in synangia (clusters); homosporous; mostly massive, fleshy ferns of tropical forests.

Family Marattiaceae. Six genera, including *Angiopteris, Marattia,* and *Danaea* with about 100 species (the extinct Carboniferous and Permian *Psaronius* belongs here).

Class Filicopsida

Leaves nonpulvinate and nonstipulate; sporangia leptosporangiate (stalked), solitary, or in sori; if soriate, then with or without indusia (protective coverings) and paraphyses (other sterile structures); mostly medium-sized sclerenchymatous (with hardened tissues) plants with cosmopolitan distribution; the class includes most common ferns.

Order Filicales

Terrestrial, epiphytic, or aquatic, homosporous ferns.

Family Osmundaceae (royal ferns). Sporangia on axes of much-reduced segments or along veins on unmodified blade, maturing nearly simultaneously, intermediate between eusporangia and leptosporangia, the annulus a lateral patch of thick-walled cells; petioles with distinct dilations; 3 genera (*Osmunda, Todea, Leptopteris*) and 20 species, plus 5 to 10 extinct genera from the Carboniferous and Permian periods.

Family Plagiogyriaceae. Leaves 1-pinnate, dimorphic, with trophophylls (sterile fronds) and sporophylls (fertile fronds), the latter contracted and containing scattered sporangia with strongly oblique annuli (rings of special cells); petiole bases swollen; stem apex and young leaves covered with mucilage from secretory hairs; 1 genus (*Plagiogyria*) with about 30 species, distributed in tropical regions.

Family Stromatopteridaceae. A strange, small, terrestrial fern with little organ differentiation; rhizome and roots merging into aboveground fronds without a sharp distinction; thought by some botanists to be related to the Psilotaceae, a family of psilotophytes, but most consider it related to the fern family Gleicheniaceae; 1 genus and species, *Stromatopteris moniliformis.*

Family Gleicheniaceae (forking ferns). Leaves mostly sprawling over other vegetation, falsely dichotomous, the segments mostly narrowly lobed; sporangia with oblique annuli and organized in simple sori; stems creeping, protostelic (its stele lacking pith and leaf gaps); *Gleichenia, Dicranopteris,* and 3 other genera with about 150 species, distributed in the tropics.

Family Matoniaceae. Leaves either fanlike, with lobed, narrow segments, or climbing, with long midribs; sporangia with oblique annuli, the simple sori covered by a thick, peltate, indusium-like structure; 2 genera (*Matonia* and *Phanerosorus*) with about 4 species, distributed in the Old World tropics.

Family Dipteridaceae (umbrella ferns). Leaves fan-shaped, venation free or finely reticulate with free, included veinlets; sporangia with oblique annuli; sori small, with paraphyses; stem with solenostele (stele with 1 leaf gap), covered with bristlelike hairs; 1 genus (*Dipteris*) with about 8 species.

Family Cheiropleuriaceae. Leaves dimorphic, with complex reticulate veins, the sterile leaves midribless, once or twice forked; sporangia with oblique annuli and 4-rowed stalks; sorus acrostichoid (covering the underside of the leaf) and with paraphyses; stem protostelic or solenostelic; 1 genus and species (*Cheiropleuria bicuspis*), distributed in the Old World tropics.

Family Polypodiaceae (true polypodies). Plants mostly epiphytic; stem scales often clathrate (latticed); leaves usually lobed or simple, with articulate (having a swollen point of separation) petioles and reticulate veins; sporangia with typical erect annuli, numerous, borne in round sori; spores bilateral, golden; gametophyte cordate (heart-shaped), nongemmiferous (not producing gemmae, tiny clusters of cells that break off and form new gametophytes); genera include *Polypodium, Pyrrosia, Platycerium,* and 30 to 50 others (depending upon the authority consulted) with 550 species.

Family Grammitidaceae (dwarf polypodies). Shaded rain forest epiphytes; leaves nonarticulate (not falling off); veins free; stem upright; stem scales with uniform walls; spores tetrahedral-globose, green; gametophyte narrowly cordate to ribbon-shaped, gemmiferous; *Grammitis* and 8 to 10 additional genera encompassing 500 species. (This family was formerly considered a subfamily of the true polypodies.)

Family Schizaeaceae. Leaves frondlike, twining, or grasslike; sporangia often borne on specialized pinnae, or leaf segments; sporangium with its annulus composed of a subapical uniserial ring of thickened cells; 6 genera (including *Anemia, Lygodium, Schizaea*) with about 160 species, mostly tropical.

Family Parkeriaceae (aquarium ferns). Plants soft, rooted in mud or floating; leaves 2- or 3-divided, upright or spreading; veins anastomosing (netted); 1 genus (*Ceratopteris*) having 4 species.

Family Platyzomataceae. Somewhat intermediate between the Schizaeaceae and Adiantaceae; 1 genus and species, *Platyzoma microphylla,* of northern Australia.

Family Adiantaceae. Sporangia in lines along the veins or on or at the vein tips, either unprotected, in grooves, covered by a more or less rolled leaf margin (called the false indusium), or rarely acrostichoid; sporangia with typical erect annuli; spores mainly tetrahedral, nonperisporial (having no outer jacket); about 60 genera containing 850 species; cosmopolitan in distribution.

Subfamily Adiantoideae (cliff brakes, maidenhair ferns, lip ferns). Mainly small and xerophytic (with adaptations to dry habitats); epipetric (growing on rocks) or terrestrial and rigid; leaves 1- to multi-pinnate; veins mostly free; *Adiantum, Cheilanthes, Pellaea, Pteris,* and 45 other genera with a total of 750 species.

Subfamily Vittarioideae (shoestring ferns). Plants leathery, epiphytic; leaves simple, pendent; veins anastomosing; *Vittaria, Antrophyum,* and 5 other genera with 100 species.

Family Lophosoriaceae. A primitive, trunkless tree fern having a short, creeping rhizome with long, golden hairs; fronds finely divided; sori round and lacking an indusium; probably related to the Metaxyaceae; 1 genus and species, *Lophosoria quadripinnata,* of high elevations in tropical America.

Family Metaxyaceae. A primitive, trunkless tree fern with a short, creeping rhizome bearing golden hairs; fronds once divided; sori round, with no indusium; probably related to the Lophosoriaceae and more distantly to *Dicksonia* in the family Cyatheaceae; 1 genus and species, *Metaxya rostrata,* of low elevations in tropical America.

Family Cyatheaceae (tree ferns). Very large, terrestrial ferns, the rhizomes trunklike and upright; indument of hairs or scales; sori indusiate, the indusium marginal and 2-lipped, or medial with indusium surrounding sorus from below; spores mostly tetrahedral, nonperisporial; genera number 7 (including *Dicksonia, Cibotium, Cyathea*) with nearly 1,000 species.

Family Dennstaedtiaceae (cup ferns, bracken). Terrestrial, medium-size to large ferns, the rhizome short to long and creeping; indument of hairs, rarely scales; sori mostly marginal with 2-lipped indusia; spores mostly tetrahedral, nonperisporial; genera number 16 (including *Dennstaedtia, Pteridium, Hypolepis, Lindsaea*) with about 400 species, mainly tropical in distribution.

Family Hymenophyllaceae (filmy ferns). Rain forest epiphytes; mostly tiny ferns with blades only 1 or 2 cells in thickness between veins; spores globose, green; gametophyte ribbon-shaped or filamentous, gemmiferous; principal genera are *Hymenophyllum* and *Trichomanes;* numerous subgenera or splinter genera are maintained by some authorities, resulting in a number of genera from 2 to 35; there are 500 species.

Family Loxsomataceae. Rare primitive ferns with sori projecting from leaf margin, borne in tubular involucra (protective coverings), and with sporangia on an elongated receptacle as in the Hymenophyllaceae, to which it probably is related; 2 genera, *Loxsoma,* with 1 species, in northern New Zealand and *Loxsomopsis,* with 3 species, from Costa Rica to Bolivia.

Family Aspleniaceae. Rhizomes mainly scaly; sori abaxial with peltate (shieldlike), reniform (kidney-shaped), oblong, or linear indusia; sporangia with typical leptosporangiate annuli; spores mostly bilateral, provided with perispore; the most common and largest family of ferns, cosmopolitan in distribution and containing about 85 genera (including *Thelypteris, Asplenium, Dryopteris, Polystichum, Elaphoglossum*) with approximately 4,000 species.

Family Blechnaceae (chain ferns). Rhizome scales nonclathrate; sori along midveins, the indusia opening inward; young foliage commonly red, turning to green; genera number 5 to 10 (including *Blechnum, Woodwardia*) with 250 species.

Family Davalliaceae. A group of largely epiphytic ferns of tropical or warm temperate regions; old fronds or leaflets fall from the plant; rhizomes scaly; sori abaxial, usually round, with a reniform or nearly tubular indusium; closely allied to the Aspleniaceae; genera number 14, including *Davallia,* the rabbit's-foot fern, and *Nephrolepis,* Boston fern relatives; about 120 species.

Order Marsileales

Aquatics with long, usually rooted rhizomes; leaves long-petioled with 2 or 4 terminal leaflets, or none; sporangia contained in sporocarps (nutlike receptacles) along petiole; heterosporous.

Family Marsileaceae. Three genera—*Marsilea* (waterclover), *Pilularia, Regnellidium*—with about 75 species.

Order Salviniales

Floating aquatics with small, rounded sporangia contained in saclike indusia; leaves sessile; heterosporous.

Family Salviniaceae (water spangles). Leaves in whorls of 3, of which 2 are oval in outline and floating, with complex hairs on the top surface, the third finely divided and hanging rootlike in the water; true roots absent; 1 genus (*Salvinia*) with 10 species.

Family Azollaceae (mosquito fern). Leaves alternate, overlapping, vertically divided into 2 lobes, the upper floating, the lower submersed; true roots present; 1 genus (*Azolla*) with 5 species.

Critical appraisal. The classification of ferns has been in a state of flux over the past several decades, but there is increasing agreement over relationships between genera. The differences in classification increasingly have to do with the level of categories deemed appropriate. Many fern groups show their relationships more by connecting links than by single character clusters shared by all members; thus many of the groups are difficult to characterize with a single description.

Relationship of ferns to "fern allies"

Some authorities still regard the ferns as being related to club mosses (Lycophyta), horsetails (Sphenophyta), and whisk ferns (Psilotophyta) and refer to all of the latter as "fern allies." In many respects, pertaining especially to their life cycle, the ferns do indeed resemble the so-called fern allies, but in other respects, especially those having to do with the vegetative body of the plant, ferns are quite distinct. (W.H.Wa./J.T.Mi.)

Lycophyta (club mosses and allies)

The Lycophyta are generally considered to constitute a division of spore-bearing vascular plants comprising the club mosses and their allies, living and fossil. Present-day lycophytes are grouped in 4 genera (some botanists divide them into 10 or more): *Lycopodium,* the club mosses or "ground pines"; *Selaginella,* the spike mosses; the unique tuberous plant *Phylloglossum;* and *Isoetes,* the quillworts. There are more than 1,000 species, widely distributed but especially numerous in the tropics. Representative extinct genera are *Lepidodendron* and *Sigillaria,* which were tree lycophytes, and *Protolepidodendron,* a herbaceous *Lycopodium*-like plant. Lycophytes are known from rocks of the Devonian Period (beginning 408 million years ago) and perhaps of the Silurian (as many as 438 million years ago). The remains of *Lepidodendron* and other extinct lycophytes form most of the great coal beds of the world.

GENERAL FEATURES

Many of the ancient lycophytes, such as *Lepidodendron,* were trees that often exceeded 30 metres (100 feet) in height. The living genera are all small plants, some erect and others low creepers. Regardless of their size or geologic age, all share certain group features. Branching is usually dichotomous; that is, the shoot tip forks repeatedly. The two branches that result may be equal in length or may be of different lengths. The leaves are generally small, although they sometimes achieved a length of one metre (three feet) in the gigantic *Lepidodendron.* Generally each leaf, or microphyll, is narrow and has an unbranched midvein, in contrast to the leaves of the ferns and seed plants, which generally have branched venation. The sporangia (spore cases) occur singly on the adaxial side (the upper side facing the stem) of the leaf. The lycophytes generally bear conelike structures called strobili, which are tight aggregations of sporophylls (sporangium-bearing leaves).

LIFE CYCLE

In the lycophytes, as in other vascular plants, there is an alternation of generations between a small, sex-cell-producing phase (gametophyte) and a conspicuous, spore-producing phase (sporophyte). The members of one of the chief living genera, *Lycopodium,* are homosporous (with just one kind of spore). They have terrestrial or subterranean gametophytes that vary in size and shape depending on the subgenera (or genera of some authors).

Underground gametophytes of *Lycopodium*

Although *Lycopodium* gametophytes are rarely found in nature, enough is known about them to recognize two fundamental types, based principally upon their mode of growth and nutrition. In some species the gametophyte becomes a small, green plant with numerous lobes, growing on the surface of the soil; the time interval between spore germination and sexual maturity of the gametophyte may be eight months to a year. In other species, including nearly all those of the north temperate zone, the gametophyte is subterranean, slower growing, and dependent upon an associated fungus for continued growth. The yellow to brown underground plant may become carrot-shaped, rod-shaped, or disk-shaped and 1 to 2 centimetres (0.4 to 0.8 inch) in length or width. Generally, a gametophyte of this type remains subterranean, and five or more years are required before it becomes sexually mature.

Gametophytes are monoecious (bisexual); *i.e.,* the sperm-producing antheridia and the egg-producing archegonia occur on the same plant. Fertilization takes place after a flagellated sperm swims to the archegonium. The embryo, or young sporophyte, consists of a shoot, a root, and a food-absorbing outgrowth called a haustorial foot. Ultimately the sporophyte becomes physiologically independent of the gametophyte, and the latter dies.

The other main extant genera—*Selaginella* and *Isoetes*—are heterosporous (having two kinds of spores). Their gametophytes are microscopic and undergo most of their development while still within the spore wall (endosporic development). Definite strobili are formed in *Selaginella,* and the sporophylls generally differ from the vegetative leaves, although not as much as in the species of *Lycopodium* that form strobili. There are two types of sporangia in *Selaginella,* called microsporangia and megasporangia; the sporophylls associated with them are termed microsporophylls and megasporophylls.

Numerous microspores are produced in the microsporangium, and cell division within the microspore wall initiates male gametophyte development. These divisions may occur before the spores are shed from the microsporangium. Final development of the male gametophyte, or microgametophyte, usually occurs on the soil prior to the release of biflagellate sperm. Usually only four large megaspores are produced in a megasporangium. Development of the female gametophyte, or megagametophyte, also may begin while the megaspore is still within the megasporangium. Free nuclear divisions (without wall formation) occur for a time, but ultimately walls appear and the megagametophyte ruptures the megaspore wall. These

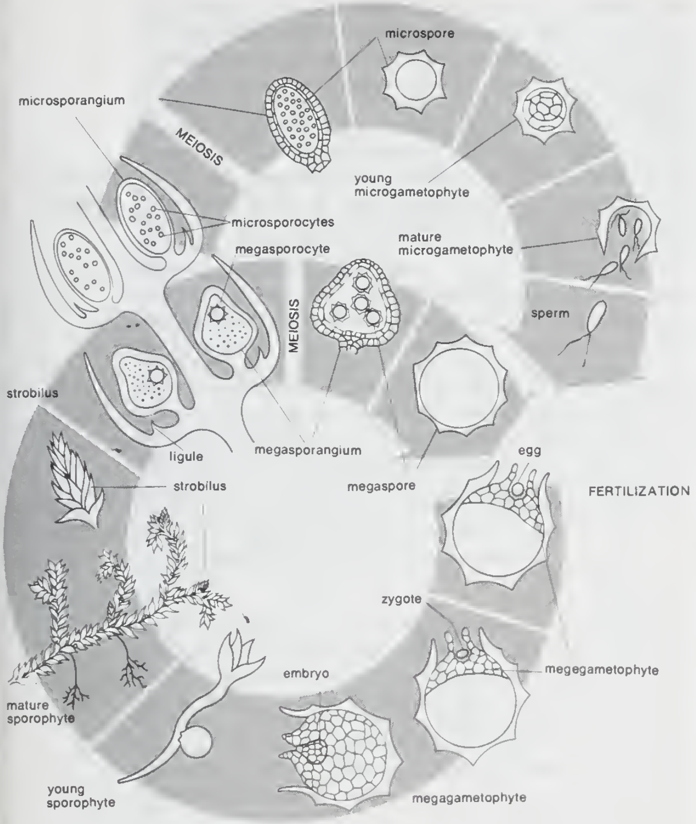


Figure 8: Life cycle of *Selaginella*.

final stages in development usually occur on the soil after the megaspore with the enclosed female gametophyte is shed from the megasporangium. Fertilization occurs when a sperm swims to an archegonium. The young sporophyte remains in physical contact with the megaspore and the enclosed female gametophyte tissue for some time.

The processes of sexual reproduction of *Isoetes* are very similar to those of *Selaginella,* except that the sperm are multiflagellate and many more spores are formed per sporangium. In fact, the microsporangia of some species are the largest among vascular plants and produce several thousand spores.

FORM AND FUNCTION

Lycopodium sporophytes

In growth habit, the sporophytes of *Lycopodium* species may rise erectly from a system of rhizomes (underground stems), or they may creep. Many are epiphytes; *i.e.,* they grow attached to tree branches or other supports. Branching is usually dichotomous, but in species with well-developed rhizomes one branch of a dichotomy usually becomes much longer and larger than the other and remains close to the surface. The shorter one may undergo several limited dichotomies, the ultimate upright branches terminating in strobili. The leaves may be spirally arranged or grouped in four vertical rows along the shoot. Each leaf has one unbranched midvein. Adventitious roots, initiated near the shoot tip, may grow within the stem cortex for some distance before emerging. The roots branch dichotomously, but no extensive root system is formed.

The stem is protostelic (without a central pith), but there is great variety in the disposition of xylem and phloem in the central vascular cylinder. Sporophylls may be aggregated into definite strobili, or there simply may be fertile and sterile regions along a stem, the sporophylls resembling vegetative leaves. Often the sporophylls of compact strobili differ from the vegetative leaves of the same plant.

Selaginella species have foliage leaves only a few millimetres long; they may be dark green or bluish and in some species are iridescent. As in *Lycopodium,* branching is usually dichotomous. The sporophyte may consist of several upright branches from a rhizome, prostrate branches creeping along the surface of the soil, or large, flat, erect, frondlike side branches from strong rhizome systems. The entire branch system often resembles a fern leaf. One distinctive feature of *Selaginella* is the rhizophore, a prop-like structure that originates at a point of branching and that forks dichotomously after making contact with the soil or a hard surface. Rhizophores are most readily seen in clambering species. Morphologically, the rhizophore is considered to be a root, although on occasion it can give rise to leafy branches if the normally leafy branches are cut off. Anisophylly (the occurrence of two sizes of leaves) occurs in most species of *Selaginella,* especially those of the wet tropics.

Ligules on leaves of *Selaginella*

Another distinctive feature in *Selaginella* is the presence of an unusual structure on the adaxial side of a leaf; this is the ligule, a peculiar tonguelike outgrowth from the leaf surface near the leaf base. Leaves of *Lycopodium* and *Selaginella* can be differentiated on this basis. The ligule, which appears very early in the development of a leaf, is a surprisingly complex structure at maturity. Its evolutionary origin is obscure. Functionally, ligules are believed to be secretory organs that, by exuding water and possibly mucilage, serve to keep young leaves and sporangia moist. Short-lived structures, they become shrunken and inconspicuous in older leaves. The ligule was a characteristic feature of the extinct giant lycophytes such as *Lepidodendron.*

Isoetes species have a plant body that is relatively small, consisting of a short axis and tufts of leaves and roots. Many species are similar in appearance to certain aquatic grasses, which are seed plants. The majority of species occur in the cooler regions of the world and are often immersed continuously in water. Each leaf is actually a sporophyll, bearing either a microsporangium or a megasporangium which is embedded in its base on the adaxial side. Each leaf also has a ligule, similar to that of *Selaginella. Isoetes* differs from both *Selaginella* and *Lycopodium* in the occurrence of secondary growth in the stem and the possession of a definite root-producing meristem. The sets of roots arise in a definite sequence, in contrast to the more or less irregularly produced roots of all other extant lower vascular plants. This sequence resembles that of its presumed ancestors *Lepidodendron* and *Pleuromeia.*

CYTOGENETICS

As in the ferns, the heterosporous representatives have much lower chromosome numbers than do the homosporous groups. Thus, *Selaginella* and *Isoetes* have $x = 9$ or 10 (*Selaginella*) and 11 (*Isoetes*), whereas *Lycopodium* and *Phylloglossum* have a wide range of higher numbers, which are correlated with sections or subgenera (or splinter genera): $x = 23$ (*Diphasiastrum*), 34 (*Lycopodium* in the strict sense), 35 (*Pseudolycopodiella*), 39 (*Lycopodiella*), 67 to 68 (*Huperzia* and *Phlegmariurus*), and 104 to 156 (*Palhinhaea*). *Phylloglossum* has $x =$ about 250. Hybridization is rare in *Selaginella* but common in *Isoetes* and the terrestrial species of *Lycopodium.*

EVOLUTION AND CLASSIFICATION

Fossil forms. The lycophytes represent a wide range of extinct and living plants that have contributed important data on evolutionary trends in primitive vascular plants. The earliest lycophytes included *Baragwanathia* and *Protolepidodendron,* dating from the early Devonian Period. Both were small herbaceous plants. During the Carboniferous Period, which followed (beginning 360 million years ago), the treelike forms of the Lepidodendrales appeared.

"Form genera"

Over the years, fossil parts of lepidodendronic plants have been discovered and assigned by taxonomists to so-called form genera, or organ genera: *Lepidophyllum* for detached leaf fossils, *Lepidostrobus* for fossil strobili. These form genera are now recognized as portions of one main fossil genus designated *Lepidodendron.* Some other lycophytes coexisting with the tree lycophytes were small herbaceous plants that resembled modern *Lycopodium* and *Selaginella* species.

Annotated classification. Groups marked with a dagger (†) in the listing below are extinct and known only from fossils.

DIVISION LYCOPHYTA (club mosses and allies)
Primitive, seedless vascular plants with true roots, stems, and leaves; sporangia associated with leaf bases, the fertile leaves often aggregated to form cones; distributed worldwide but concentrated in the tropics.

†Order Protolepidodendrales
Extinct herbaceous (rarely woody), homosporous lycophytes; about 8 genera, including *Baragwanathia* and *Protolepidodendron.*

†Order Lepidodendrales
Extinct tree lycophytes, therefore capable of secondary growth; heterosporous, with some strobili (cones) forming seedlike structures; about 6 genera, including *Lepidodendron* and *Sigillaria.*

Order Lycopodiales
Living and extinct plants with primary growth only; homosporous; 2 living genera, *Lycopodium,* with 200 species, mostly tropical, and *Phylloglossum,* with 1 species, restricted to Australia and New Zealand; includes the extinct *Lycopodites.*

Order Selaginellales
Living and extinct plants with primary growth only; heterosporous; the sole living genus is *Selaginella,* with nearly 800 species, widely distributed around the world; *Selaginellites* is an extinct genus.

Order Isoetales
Living and extinct plants with secondary growth; heterosporous, with endosporic gametophytes; *Isoetites* is an extinct genus; a specialized group of species from the high Andes Mountains is sometimes segregated as a distinct genus, *Stylites;* for many years the species of *Isoetes* were difficult to distinguish, but, since the discovery that frequent hybridization was obscuring the differences between species, they are more clearly understood; *Isoetes* includes more than 100 species in swampy, cooler parts of the world.

†Order Pleuromeiales
Extinct unbranched plants, with subterranean, rootlike rhizophores; heterosporous; a single fossil genus, *Pleuromeia.*

Critical appraisal. This group is now generally recognized as a division; in the past, however, it was commonly treated as a class, Lycopsida. Students of the Lycophyta are finding increasing evidence to support the division of *Lycopodium* and *Selaginella* each into two or more genera. *Lycopodium* has 3 to 11 groups that might be recognized as distinct genera, based on different chromosome numbers, spore sculpturing, and gametophyte morphology. Similarly, *Selaginella* is divided into two or three groups on the basis of differences in spores and leaves (those with two kinds of leaves and those with leaves all alike). The groupings appear to be natural, but it is too soon to say whether these subdivisions will receive general acceptance as genera among botanists.

Sphenophyta (horsetails)

The Sphenophyta, or Equisetophyta, are a division of primitive spore-bearing vascular plants. Most members of the group are extinct and known only from their fossilized remains. The sole living genus, *Equisetum,* is made up of about 15 species of very ancient herbaceous plants, the horsetails and scouring rushes. Extinct members of the division, some of which have been traced back as far as the Devonian Period (408 to 360 million years ago), include many herbaceous Equisetales, shrubby Hyeniales, vinelike Sphenophyllales, and trees of the family Calamitaceae.

GENERAL FEATURES

Sphenophytes, fossil and living, characteristically have whorled leaves and branches and conspicuously jointed stems, which in many cases are also ribbed. Reproductive structures are present in the form of greatly compressed stems called cones, or strobili, which form at the ends of branches.

The giant extinct horsetails (*Calamites*) were trees up to 1 metre (3 feet) in diameter and 30 metres (100 feet) in height. Their leaves—like those of extant horsetails—were arranged in spokelike whorls at regular intervals along the jointed stems. In the Sphenophyllales, an extinct order of scrambling sphenophytes, the leaves were wedge-shaped, with a repeatedly forking (dichotomous) venation system (sphenophylls). The order Hyeniales included shrublike plants with inconspicuous leaves arranged in rather indistinct whorls.

The living species of *Equisetum* are widely distributed in the Northern Hemisphere. Most of them are less than one metre (three feet) tall. There are reports of specimens of *E. giganteum,* from the American tropics, that attain a height of about 10 metres with a stem diameter of only 4 centimetres (1.6 inches); support is apparently provided by their habit of growing in dense stands in their natural environment. The majority of *Equisetum* species are found in wet or damp habitats, often in shaded locations along streams, ditches, and canals; some species, however, have become adapted to drier and sunnier conditions.

Equisetum habitats

The extant sphenophytes have had little economic importance. The extinct giant types contributed to the coal beds formed in the Carboniferous Period (360 to 286 million years ago). Living horsetails have been used as scouring agents, their cleansing value being attributed to the abrasive action of the silica-laden walls of certain of their cells. Silica is only one of several minerals that horsetails selectively accumulate in their bodies. Gold is another—up to 0.15 gram per kilogram (4.5 ounces per ton) of plants—not economically feasible to mine but a certain indication of the availability of such ore deposits in the soil. Horses foraging on stands of *Equisetum* have been known to die from severe intestinal inflammation.

LIFE CYCLE

Horsetails, like other vascular plants, display an alternation of generations: an asexual phase, represented by a sporophyte (the horsetail plant), and a sexual phase, the gametophyte, an inconspicuous, delicate, green plant. Each year, many gametophytes are initiated from spores, but apparently very few produce sporophytes in nature. Horsetails apparently survive mainly by vegetative reproduction rather than by a regular dependence on the sexual cycle.

Some horsetails carry terminal cones (strobili) on green aerial branches. Other species, however, have separate upright, aerial branches for vegetative and for reproductive shoots. In these species the strobilate branches appear first, and, after the spores are shed, the green vegetative shoots develop. The fertile components of the strobilus are called sporangiophores; each consists of a stalk bearing a flattened disk at its apex, on the lower edge of which is a ring of 5 to 10 sporangia, each one opening and shedding spores by a longitudinal slit on its inner side. The Carboniferous treelike horsetails and their smaller allies are believed to have possessed the most elaborate reproductive strobili known among the vascular plants.

Sphenophytes are homosporous, producing only one kind of spore. The spores have four bands, or elaters, which coil and uncoil in response to changes in humidity, assisting in

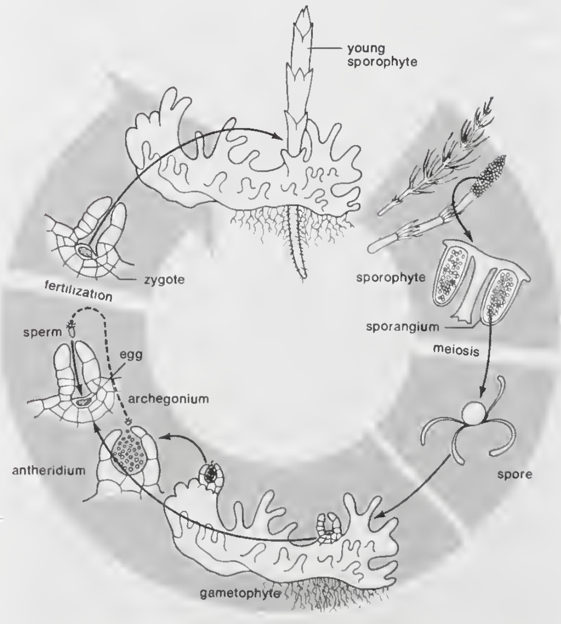


Figure 9: Life cycle of *Equisetum.*

the dispersal of the spores. Under low light intensity and high humidity, the spores germinate to form small, flattened, green gametophytes. After a period of development, these gametophytes come to resemble miniature green pincushions up to 3 centimetres (1.2 inches) in diameter. Eggs are produced in archegonia, at the bases of upright lobes on the gametophytes, and sperm are produced in antheridia, present on the lobes. The egg is fertilized in the archegonium by a sperm, forming a zygote which, by continued divisions, develops an embryo within the archegonium. The embryo (young sporophyte) is nourished by the gametophyte until it develops its own shoot and roots. One gametophyte may support two or more young sporophytes before it ultimately dies and decays.

FORM AND FUNCTION

Jointed structure of stem

The sporophyte of a typical sphenophyte consists of stem, leaves, and roots. The underground part of the stem (the rhizome) and the aerial part show the same basic organization. They consist of distinct segments united end to end at the nodes, which are the origins of the roots and leaves. (This jointed structure is the source of the alternative name Articulatae, which was applied to the Sphenophyta by some earlier authorities.)

The slender, herbaceous stems of *Equisetum* have hollow internodes. The whorled leaves are greatly reduced and nonphotosynthetic and are united laterally at each node to form a toothed sheath around the stem. Each leaf has a single, unbranched midvein. Secondary growth, by which girth increases annually, was characteristic of the extinct Calamitaceae and Sphenophyllales. In *Equisetum* the vascular strands are small and round, surrounding a large pith cavity. In the majority of species, the cell walls of the outer cell layer (epidermis) are thick and contain silica deposits. Branch buds are initiated at the nodes, but their subsequent development is dependent on the growth characteristics of the particular species. In relatively unbranched species, bud growth remains inhibited.

Spores develop in spore cases (sporangia), which are borne on sporangiophores. These are organized into strobili, which may be associated with sterile bracts (much reduced leaves)—as in Sphenophyllales and Calamitaceae—or may be without them (Hyeniales, Equisetaceae). Each sporangiophore in *Equisetum* has 5 to 10 sporangia. The entire sporangiophore may have arisen as a condensed, dichotomous branch system, with each sporangium occupying the end of a branch but lying parallel to the stalk of the sporangiophore.

CYTOGENETICS

Chromosome numbers in *Equisetum* are uniformly $x = 108$. Several hybrids are known, but all are sterile, since there is no doubling of the chromosome number to allow chromosome pairing and consequent production of viable spores.

EVOLUTION AND CLASSIFICATION

Evolutionary development. Certain Sphenophyta flourished as trees (*e.g., Calamites* species) during the coal-forming Carboniferous Period, but the earliest sphenophytes appeared as early as the Devonian. In its fossil history the division constituted a much larger portion of the flora of the Earth than it does at the present time.

Equisetum, which may also have been present during the Carboniferous, is perhaps one of the oldest living genera of vascular plants. The more primitive species have perennial, green shoots. The advanced species have annual, green, branched, vegetative shoots and often nongreen, unbranched, fertile shoots. Intermediate combinations of these features occur in some species.

Annotated classification. Botanists recognize four to six different orders in the division. Only one order, Equisetales, has both living and extinct species; all others comprise extinct sphenophytes. The latter are indicated by a dagger (†) in the listing below.

DIVISION SPHENOPHYTA (EQUISETOPHYTA; horsetails)
Extinct and living primitive, seedless, homosporous vascular plants with jointed, ribbed stems and whorls of leaves at regular intervals along the stem.

†Order Hyeniales (Protoarticulatae)
Extinct shrublike plants, with short, forked leaves in whorls; 1 family: Hyeniaceae (now placed with the Filicophyta—true ferns—by some paleobotanists).

†Order Pseudoborniales
One family, Pseudoborniaceae, with a single extinct species, *Pseudobornia ursina;* 15 to 20 metres (50 to 65 feet) tall.

†Order Sphenophyllales
Extinct scrambling or vinelike understory plants, 1 metre (3 feet) tall, with small, wedge-shape leaves; 2 families: Sphenophyllaceae and Cheirostrobaceae.

Order Equisetales
Two families: Calamitaceae, extinct tree horsetails; and Equisetaceae, herbaceous living horsetails and fossil allies with needlelike leaves in whorls along the stem; about 15 extant species in the genus *Equisetum* and several extinct species in the genus *Equisetites; Equisetum* is distributed predominantly in the Northern Hemisphere.

Critical appraisal. The extant genus *Equisetum* is a small remnant of a once diverse and dominant plant group. Although the genus includes two rather distinct groups, modern botanists recognize but a single genus.

Psilotophyta (whisk ferns)

The Psilotophyta are a division of spore-bearing vascular plants with the simplest structure known of any extant group. Psilotophytes have no true roots or leaves.

GENERAL FEATURES

Geographic distribution

The two genera, *Psilotum* (whisk ferns) and *Tmesipteris,* are terrestrial or epiphytic (living on other plant surfaces). *Psilotum,* an upright, green plant that looks like a leafless shrub about 30 centimetres (12 inches) high, is tropical and subtropical in distribution, reaching as far north as Florida. *Tmesipteris,* a hanging green epiphyte, is found in Australia, New Caledonia, New Zealand, the Philippines, and other islands of the South Pacific area. *Psilotum* is sometimes grown in greenhouses or even as a houseplant, but *Tmesipteris* has never been satisfactorily cultivated.

The conspicuous green plant, the sporophyte, consists of an aerial system and an underground stem (rhizome) system, both of which repeatedly fork dichotomously—*i.e.,* into two equal branches. No roots are present, the rhizome performing the functions of a root. In *Psilotum* the "leaves" are small, scalelike outgrowths without vascular tissue; in *Tmesipteris* they are larger, flattened structures, each with one unbranched midvein.

LIFE CYCLE

Psilotophytes alternate between two forms: gametophytes, which reproduce sexually, and sporophytes, which reproduce asexually. The psilotophytes are homosporous, the sporangia (spore cases) producing a single type of spore. After the spores are released by longitudinal splitting of the sporangium, they germinate to form the sex-cell-bearing plants—gametophytes. These plants are very small, measuring about 1 millimetre (0.04 inch) in diameter and a few millimetres in length. They may grow on trunks of trees or underground. Like the sporophyte, the gametophyte branches repeatedly. It is devoid of chlorophyll, however, and lives a saprophytic existence (*i.e.,* obtaining food from decaying organic matter), presumably aided by a fungus living within it.

The gametophytes are so similar to the underground rhizomes that they can be identified only by their gametangia—the sperm-producing antheridia and the egg-producing archegonia—which are scattered over the surface and are intermingled.

The first division of the zygote (fertilized egg) produces two cells. The outer cell continues to divide and forms the axis of the developing sporophyte. The inner cell gives rise to a multicellular "foot," which is in close contact with the gametophyte and presumably functions in the transfer of nutrients from the gametophyte to the young sporophyte.

FORM AND FUNCTION

Psilotophytes have a simple internal organization. In the rhizomes of both genera, the vascular tissue is in the form of a slender cylinder (protostele) occupying the centre of the stem. The upper branches of *Psilotum* are also protostelic. In the lower branches of *Psilotum* and in the stems of *Tmesipteris,* the vascular tissue forms a siphonostele, a cylinder surrounding nonvascular cells. The cells of the outer layers of the stem contain chloroplasts, and in *Psilotum* this region accounts for most of the photosynthesis.

Sporangia are three-lobed in *Psilotum* and two-lobed in *Tmesipteris.* The sporangia in *Psilotum* appear generally to be located in the axil of scalelike "leaves" but are actually on very short branches.

CYTOGENETICS

Chromosome numbers in the psilotophytes range from $x = 52$ to $4x = 208$. Hybridization is rare, the only reported hybrid (from Hawaii) being between the two species of *Psilotum.*

CLASSIFICATION

Annotated classification.

DIVISION PSILOTOPHYTA (whisk ferns)

The simplest of the primitive, seedless, homosporous plants, lacking roots; the upright stem or flattened "leaves" serve as photosynthetic organs and the horizontal stem serves to anchor the plant and to absorb nutrients.

Order Psilotales

One family, Psilotaceae, with 2 genera: *Psilotum,* a leafless, shrublike plant of the tropics and subtropics, with 2 species; and *Tmesipteris,* a hanging epiphyte with flattened leaflike structures, found in Oceania and the South Pacific area, with 7 species.

Critical appraisal. For many years this group was thought to be primitive in its simplicity and related to simple plants of Silurian and Devonian times (438 to 360 million years ago). However, the lack of an intervening fossil record and differences in morphology have led botanists to conclude that the Psilotophyta are unrelated to those early vascular plants. Whether the psilotophytes represent reduced lycophytes or are related to the ferns, such as the relatively undifferentiated *Stromatopteris,* is still in question. (E.M.G./J.T.Mi.)

BIBLIOGRAPHY. F.O. BOWER, *The Ferns (Filicales): Treated Comparatively with a View to Their Natural Classification,* vol. 1, *Analytical Examination of the Criteria of Comparison,* vol. 2, *The Eusporangiatae and Other Relatively Primitive Ferns,* and vol. 3, *The Leptosporangiate Ferns* (1923–28), is a classic work of comparative morphology and systematics that emphasizes the need, now being realized, for a broad spectrum of comparative data. A comprehensive summary of paleobotanical knowledge is provided in THOMAS N. TAYLOR, *Paleobotany: An Introduction to Fossil Plant Biology* (1981). The American Fern Society and the British Pteridological Society assemble the record of current research in the field in their publications *American Fern Journal* (quarterly), *Fiddlehead Forum* (bimonthly), *The Fern Gazette* (annual), and *Pteridologist* (annual).

The abundance and diversity of pteridophytes are the focus of HERMANN CHRIST, *Die Geographie der Farne* (1910), still an important broad treatment of fern distribution; JOHN T. MICKEL, *How to Know the Ferns and Fern Allies* (1979), the first manual to cover all of North America, with keys, brief descriptions, and illustrations; ROLLA M. TRYON and ALICE F. TRYON, *Ferns and Allied Plants* (1982), a good summary of the genera of tropical American pteridophytes with descriptions, maps, discussions, and many illustrations; JOHN T. MICKEL and JOSEPH M. BEITEL, *Pteridophyte Flora of Oaxaca, Mexico* (1988), the best illustrated and most comprehensive pteridophyte manual for Latin America; and R.E. HOLTTUM, *A Revised Flora of Malaya: An Illustrated Systematic Account of the Malayan Flora, Including Commonly Cultivated Plants,* vol. 2, *Ferns of Malaya* (1954), a well-illustrated enumeration and description of ferns that presents many of the author's ideas of systematic relationship.

Life cycle and habitats are discussed in A.F. DYER, *The Experimental Biology of Ferns* (1979), a series of essays on ecology, cytogenetics, reproduction, chemistry, and development; A.F. DYER and CHRISTOPHER N. PAGE (eds.), *Biology of Pteridophytes* (1985), a collection of symposium papers on a broad range of topics; F. GORDON FOSTER, *Ferns to Know and Grow,* 3rd rev. ed. (1984), a well-known book of horticulture with many helpful tips on cultivation; BARBARA JOE HOSHIZAKI, *Fern Growers Manual* (1975), a good introduction to horticulture with encyclopaedic information on the species in cultivation; and CHRISTOPHER N. PAGE, *Ferns: Their Habitats in the British and Irish Landscape* (1988), with excellent illustrations of habitats and ecology.

Studies of form and function include K.R. SPORNE, *The Morphology of Pteridophytes: The Structure of Ferns and Allied Plants,* 4th ed. (1975), a concise summary of ideas on fern structure; B.K. NAYAR and S. KAUR, "Gametophytes of Homosporous Ferns," *The Botanical Review* 37:295–396 (1971), a thorough summation of the knowledge of the haploid generation of ferns, with an extensive bibliography; JOHN T. MICKEL, *The Home Gardener's Book of Ferns* (1979), a useful compilation of information on fern morphology, diversity, and cultivation; and LENORE W. MAY, "The Economic Uses and Associated Folklore of Ferns and Fern Allies," *The Botanical Review* 44:491–528 (1978), a summary of the diverse uses to which ferns have been put.

For the origin and evolution of ferns and fern allies, see I. MANTON, *Problems of Cytology and Evolution in the Pteridophyta* (1950), a milestone in the biology of ferns containing, for the first time, accurate data on chromosomes in relation to evolution and systematics; RICHARD A. WHITE (ed.), "Taxonomic and Morphological Relationships of the Psilotaceae: A Symposium," *Brittonia* 29:1–68 (1977), a series of papers on structure, relationships, and fossil history; and J.D. LOVIS, "Evolutionary Patterns and Processes in Ferns," *Advances in Botanical Research* 4:229–439 (1977), an outstanding summary of the knowledge of fern phylogeny and classification. Also see appropriate sections of ROBERT F. SCAGEL *et al., An Evolutionary Survey of the Plant Kingdom* (1965); ERNEST M. GIFFORD and ADRIANCE S. FOSTER, *Morphology and Evolution of Vascular Plants,* 3rd ed. (1989); and DAVID W. BIERHORST, *Morphology of Vascular Plants* (1971), which provides detailed treatments of vascular plants together with theory and interpretation.

Nomenclature for the taxonomy of pteridophytes is provided in EDWIN BINGHAM COPELAND, *Genera Filicum: The Genera of Ferns* (1947), a valuable treatment of the classification and characteristics of ferns, containing many of the author's original correlations. Other works on classification include R.L. HAUKE, "The Taxonomy of Equisetum: An Overview," *New Botanist* 1:89–95 (1974); J.A. CRABBE, A.C. JERMY, and JOHN T. MICKEL, "A New Generic Sequence for the Pteridophyte Herbarium," *The Fern Gazette* 11:141–162 (1975), a list of pteridophyte genera in a phylogenetic sequence; and BENJAMIN ØLLGAARD, "A Revised Classification of the Lycopodiaceae s. lat.," *Opera Botanica* 92:153–178 (1987), a clear, detailed discussion of the taxonomic characters, genera, and species groups of the family, and *Index of the Lycopodiaceae* (1989), a listing of all the names, references, and type (original) specimens.

(W.H.Wa./E.M.G./J.T.Mi.)

Finland

An independent republic in northern Europe, Finland (Finnish: Suomi; in full Suomen Tasavalta, or Republic of Finland; Swedish: Finland, or Republiken Finland) is one of the world's most northern and geographically remote countries and is subject to a severe climate. It is bordered on the north by Norway, on the northwest by Sweden, on the southwest by the Gulf of Bothnia, on the south by the Gulf of Finland, and on the east by Russia. Its area is 130,559 square miles (338,145 square kilometres), of which the Åland Islands, an archipelago at the entrance to the Gulf of Bothnia, constitute 590 square miles. About one-third of the territory of Finland—most of the *lääni* (province) of Lapland—lies north of the Arctic Circle. The capital is Helsinki.

Finland was part of Sweden from the 12th century until 1809. It then was a Russian grand duchy until, following the Russian Revolution, the Finns declared independence on Dec. 6, 1917. Finland's area decreased by about one-tenth during the 1940s, when it ceded the Petsamo (Pechenga) area, which had been a corridor to the ice-free Arctic coast, and a large part of southeastern Karelia to the Soviet Union (ceded portions now in Russia).

Throughout the Cold War era, Finland maintained a carefully neutral political position, although a 1948 treaty with the Soviet Union (terminated 1991) required Finland to repel any attack on the Soviet Union carried out through Finnish territory by Germany or any of its allies. Since World War II, Finland has steadily increased its trading and cultural relations with other countries. Under a U.S.-Soviet agreement, Finland was admitted to the United Nations in 1955. Since 1955, Finland has sent representatives to the Nordic Council, which makes suggestions to member countries on the coordination of policies.

Finland's international activities became more widely known when the Conference on Security and Cooperation in Europe, which resulted in the creation of the Helsinki Accords, was held in that city in 1975. Finland has continued to have especially close ties with the other Scandinavian countries, sharing a free labour market and participating in various economic, cultural, and scientific projects. Finland became a member of the European Union (and its constituent European Community) in 1995.

This article is divided into the following sections:

Physical and human geography

THE LAND

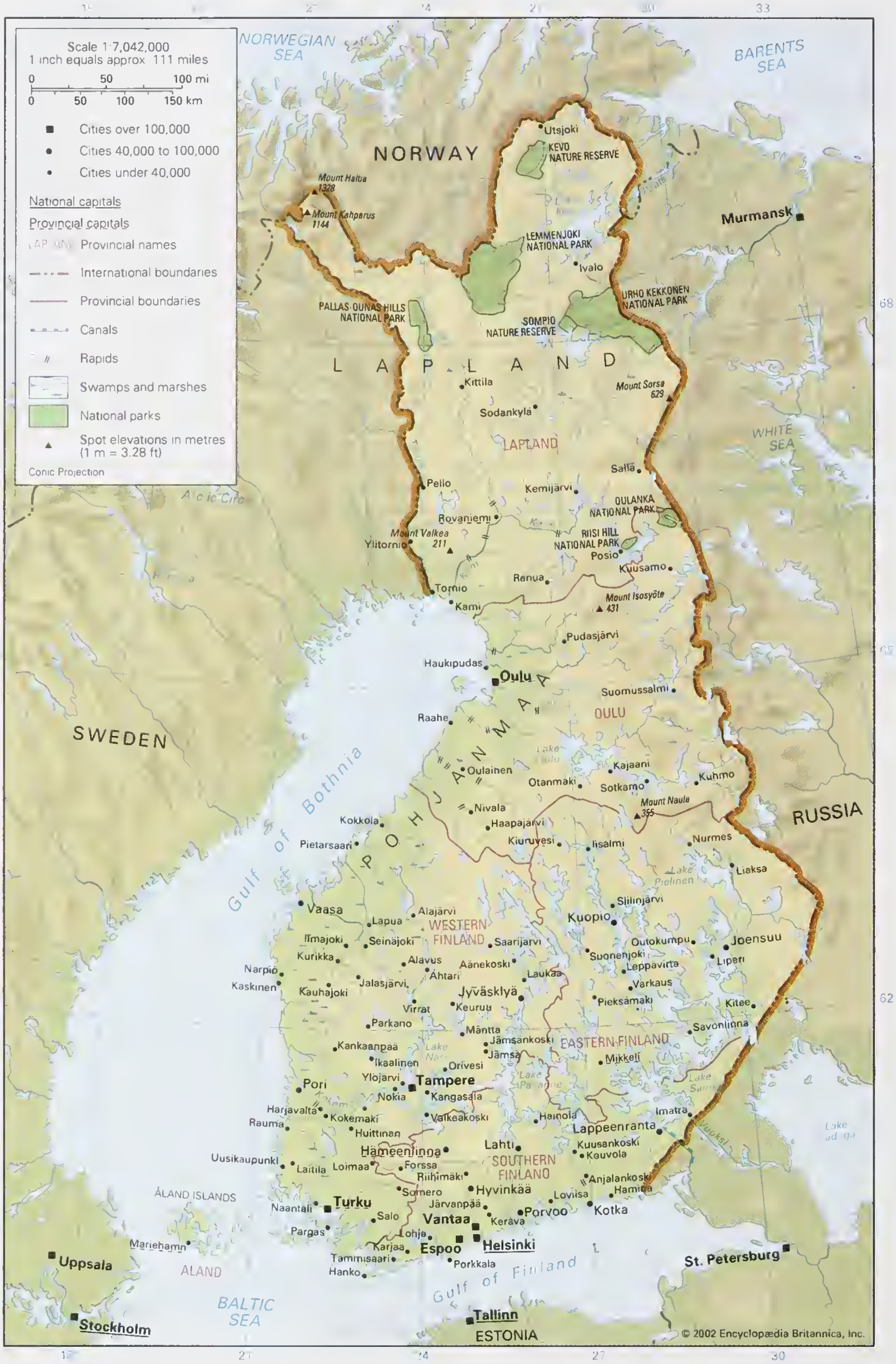
Relief. Finland is heavily forested and contains some 55,000 lakes, numerous rivers, and extensive areas of marshland; viewed from the air, Finland looks like an intricate blue and green jigsaw puzzle. Except in the northwest, relief features do not vary greatly, and travelers on the ground or on the water can rarely see beyond the trees in their immediate vicinity. The landscape nevertheless possesses a striking—if sometimes bleak—beauty.

Ancient underlying rocks

Finland's underlying structure is a huge worn-down shield composed of ancient rock, mainly granite, dating from Precambrian time (3,800,000,000 to 540,000,000 years ago). The land is low-lying in the southern part of the country and higher in the centre and the northeast, while the few mountainous regions are in the extreme northwest, adjacent to Finland's borders with Sweden and Norway. In this area there are several high peaks, including Mount Haltia, which, at 4,357 feet (1,328 metres), is Finland's highest mountain.

The coastline of Finland, some 2,760 miles (4,600 kilometres) in length, is extremely indented and dotted with thousands of islands. The greatest number of these are to be found in the southwest, in the Turun (Turku) archipelago, which merges with the Åland Islands in the west. The southern islands in the Gulf of Finland are mainly of low elevation, while those lying along the southwest coastline may rise to heights of more than 400 feet.

The relief of Finland was greatly affected by the Ice Age. The retreating continental glacier left the bedrock littered with morainic deposits in formations of eskers, remarkable winding ridges of stratified gravel and sand, running northwest to southeast. One of the biggest formations is the Salpausselkä ridges, three parallel ridges running across southern Finland in an arc pattern. The weight of the gla-



ciers, sometimes miles thick, depressed the Earth's crust by many hundreds of feet. As a consequence, areas that have been released from the weight of the ice sheets have risen and continue to rise, and Finland is still emerging from the sea. Land rise of some 0.4 inch (9 millimetres) annually in the narrow part of the Gulf of Bothnia is gradually turning the old sea bottom into dry land.

Drainage and soils. Finland's inland waters occupy almost 10 percent of the country's total area; there are 10 lakes of more than 100 square miles in area and tens of thousands of smaller ones. The largest lake, Saimaa, in the southeast, covers about 1,700 square miles. There are many other large lakes near it, including Päijänne and Pielinen, while Oulu is near Kajaani in central Finland, and Inari is in the extreme north. Away from coastal regions, many of Finland's rivers flow into the lakes, which are generally shallow—only three lakes are deeper than about 300 feet. Saimaa itself drains into the much larger Lake Ladoga in Russian territory via the Vuoksi (Vuoksa) River. Drainage from Finland's eastern uplands is through the lake system of Russian Karelia to the White Sea.

In the extreme north, the Paats River and its tributaries drain large areas into the Arctic. On the western coast, a series of rivers flow into the Gulf of Bothnia, including the

MAP INDEX

Political subdivisions

Åland (Ahvenamaa)	60 15 N 20 00 E
Eastern Finland (Itä-Suomi)	62 30 N 28 00 E
Lapland	68 00 N 27 00 E
Oulu	65 00 N 27 00 E
Southern Finland (Etelä-Suomi)	60 50 N 26 00 E
Westem Finland (Länsi-Suomi)	62 00 N 24 00 E

Cities and towns

Äänekoski	62 36 N 25 44 E
Åbo, see Turku	
Ähtari	62 34 N 24 06 E
Alajärvi	63 00 N 23 49 E
Alavus	62 35 N 23 37 E
Anjalankoski	60 44 N 26 47 E
Espoo (Esbo)	60 13 N 24 40 E
Forssa	60 49 N 23 38 E
Haapajärvi	63 45 N 25 20 E
Hämeenlinna (Tavastehus)	61 00 N 24 27 E
Hamina (Fredrikshamn)	60 34 N 27 10 E
Hanko	59 50 N 22 57 E
Harjavalta	61 19 N 22 08 E
Haukipudas	65 11 N 25 21 E
Heinola	61 13 N 26 02 E
Helsinki	60 10 N 24 58 E
Huittinen	61 11 N 22 42 E
Hyvinkää	60 38 N 24 52 E
Iisalmi	63 34 N 27 11 E
Ikaalinen	61 46 N 23 03 E
Ilmajoki	62 44 N 22 34 E
Imatra	61 10 N 28 46 E
Ivalo	68 39 N 27 36 E
Jalasjärvi	62 30 N 22 45 E
Jämsä	61 52 N 25 12 E
Jamsänkoski	61 55 N 25 11 E
Jarvenpaa	60 28 N 25 06 E
Joensuu	62 36 N 29 46 E
Jyväskylä	62 14 N 25 44 E
Kajaani	64 14 N 27 41 E
Kangasala	61 28 N 24 05 E
Kankaanpää	61 48 N 22 25 E
Karjaa (Karis)	60 05 N 23 40 E
Kaskinen	62 23 N 21 13 E
Kauhajoki	62 26 N 22 11 E
Kemi	65 44 N 24 34 E
Kemijarvi	66 40 N 27 25 E
Kerava	60 24 N 25 07 E
Keuruu	62 16 N 24 42 E
Kitce	62 06 N 30 09 E
Kittilä	67 40 N 24 54 E
Kiuruvesi	63 39 N 26 37 E
Kokemaki	61 15 N 22 21 E
Kokkola (Karleby)	63 50 N 23 07 E
Kotka	60 28 N 26 55 E
Kouvola	60 52 N 26 42 E
Kuhmo	64 08 N 29 31 E
Kuopio	62 54 N 27 41 E
Kurikka	62 37 N 22 25 E
Kuusamo	65 58 N 29 11 E
Kuusankoski	60 54 N 26 38 E
Lahti	60 58 N 25 40 E
Laitila	60 55 N 21 45 E
Lappeenranta (Villmanstrand)	61 04 N 28 11 E
Lapua	62 57 N 23 00 E
Laukaa	62 25 N 25 57 E
Leppävirta	62 29 N 27 47 E
Lieksa	63 19 N 30 01 E
Liperi	62 32 N 29 22 E
Lohja	60 15 N 24 05 E
Loimaa	60 51 N 23 03 E
Loviisa	60 27 N 26 14 E
Mänttä	62 02 N 24 38 E
Mariehamn (Maarianhamina)	60 06 N 19 57 E
Mikkeli (Sankt Michel)	61 41 N 27 15 E
Naantali	60 27 N 22 02 E
Narpiö (Narpes)	62 28 N 21 20 E
Nivala	63 55 N 24 58 E
Nokia	61 28 N 23 30 E
Nurmes	63 33 N 29 07 E
Orivesi	61 41 N 24 21 E
Otanmaki	64 07 N 27 06 E
Oulainen	64 16 N 24 48 E
Oulu (Uleåborg)	65 01 N 25 28 E
Outokumpu	62 44 N 29 01 E
Pargas	60 18 N 22 18 E
Parkano	62 01 N 23 01 E
Pello	66 47 N 23 55 E
Pieksämäki	62 18 N 27 08 E
Pietarsaari	63 40 N 22 42 E
Pori (Bjorneborg)	61 29 N 21 47 E
Porkkala	59 59 N 24 26 E
Porvoo (Borgå)	60 24 N 25 40 E
Posio	66 06 N 28 09 E
Pudasjärvi	65 23 N 26 55 E
Raahe	64 41 N 24 29 E
Ranua	65 55 N 26 32 E
Rauma	61 08 N 21 30 E
Riihimäki	60 45 N 24 46 E
Rovaniemi	66 30 N 25 43 E
Saarijärvi	62 38 N 23 19 E
Salla	66 50 N 28 40 E
Salo	60 23 N 23 08 E
Savonlinna	61 52 N 28 53 E
Seinäjoki	62 47 N 22 50 E
Siilinjärvi	63 05 N 27 40 E
Sodankylä	67 25 N 26 36 E
Somero	60 37 N 23 32 E
Sotkamo	64 08 N 28 25 E
Suomussalmi	64 53 N 29 05 E
Suonenjoki	62 37 N 27 08 E
Tammisaari (Ekenäs)	59 58 N 23 26 E
Tampere (Tammerfors)	61 30 N 23 45 E
Tornio	65 51 N 24 08 E
Turku (Åbo)	60 27 N 22 17 E
Utsjoki	69 53 N 27 00 E
Uusikaupunki	60 48 N 21 25 E
Vaasa (Vasa)	63 06 N 21 36 E
Valkeakoski	61 16 N 24 02 E
Vammala	60 56 N 22 54 E
Vantaa (Vanda)	60 18 N 24 51 E
Varkaus	62 19 N 27 55 E
Virrat	62 14 N 23 47 E
Ylitornio	66 18 N 23 40 E
Ylojarvi	61 33 N 23 36 E

Physical features and points of interest

Åland Islands (Ahvenanmaa)	60 15 N 20 00 E
Baltic Sea	59 00 N 21 00 E
Barents Sea	71 00 N 33 00 E
Bothnia, Gulf of	63 00 N 20 00 E
Finland, Gulf of	60 00 N 27 00 E
Haltia, Mount (Haldefjäll)	69 18 N 21 16 E
Inari, Lake	69 00 N 28 00 E
Isosyöte, Mount	65 37 N 27 35 E
Kahperus, Mount	69 07 N 21 13 E
Kemi, *river*	65 47 N 24 30 E
Kevo Nature Reserve	69 30 N 26 30 E
Kokemäen, *river*	61 33 N 21 42 E
Lapland (Lappland), *region*	68 00 N 25 00 E
Lemmenjoki National Park	69 00 N 25 30 E
Näsi, Lake	61 37 N 23 42 E
Naula, Mount	63 53 N 28 13 E
Norwegian Sea	71 00 N 21 00 E
Oulanka National Park	66 25 N 29 15 E
Oulu, *river*	65 01 N 25 25 E
Oulu, Lake	64 20 N 27 15 E
Paats, *river*	69 43 N 30 06 E
Päijänne, Lake	61 35 N 25 30 E
Pallas-Ounas Hills National Park	68 00 N 24 00 E
Pielinen, Lake	63 15 N 29 40 E
Pohjanmaa, *lowland*	64 00 N 25 00 E
Riisi Hill National Park	66 15 N 28 20 E
Saimaa, Lake	61 15 N 28 15 E
Sompio Nature Reserve	68 10 N 27 15 E
Sorsa, Mount	67 24 N 29 38 E
Tornio, *river*	65 48 N 24 08 E
Urho Kekkonen National Park	68 10 N 28 30 E
Valkea, Mount	66 14 N 24 39 E
Vuoksi, *river*	61 03 N 30 11 E

Tornio, which forms part of Finland's border with Sweden, and the Kemi, at 343 miles Finland's longest river. In the southwest the Kokemäen, one of Finland's largest rivers, flows out past the city of Pori (Björneborg). Other rivers flow southward into the Gulf of Finland.

Glacial influences in soil formation

Soils include those of the gravelly type found in the eskers, as well as extensive marine and lake postglacial deposits in the form of clays and silts, which provide the country's most fertile soils. The northern third of Finland has thick layers of peat. Such marshland still covers as much as 30 percent of the countryside. In the Åland Islands the soils are mainly clay and sand.

Climate. The part of Finland north of the Arctic Circle suffers extremely severe and prolonged winters. Temperatures can fall as low as −22° F (−30° C). In these latitudes the snow never melts from the north-facing mountain slopes, but in the short summer (Lapland has about two months of the midnight sun), from May to July, temperatures can reach as high as 80° F (27° C). Farther south the temperature extremes are slightly less marked. Annual precipitation, about one-third of which falls as sleet or snow, is about 25 inches in the south and a little less in the north. All Finnish waters are subject to some surface freezing during the winter.

Plant and animal life. Much of Finland is dominated by conifers, but in the extreme south there is a zone of deciduous trees comprising mainly birch, hazel, aspen, maple, elm, linden, and alder. The conifers are mainly pine and spruce. Pine extends to the extreme north, where it can be found among the dwarf arctic birch and pygmy willow. Lichens become increasingly common and varied in kind toward the north. In autumn the woods are rich in edible fungi. More than 1,000 species of flowering plants have been recorded. The sphagnum swamps, which are widespread in the northern tundra or bogland area, yield harvests of cloudberries, as well as plagues of mosquitoes.

Finland is relatively rich in wildlife. Seabirds, such as the black-backed gull and the arctic tern, nest in great numbers on the coastal islands; waterfowl, such as the black and white velvet scotcr duck, nest on inland lakes. Other birds include the Siberian jay, picd wagtail, and, in thc north, the eagle. Many birds migrate southward in winter. Native woodland animals include bear, clk, wolf, wolvcrinc, lynx, and Finnish clk. Wild reindccr havc almost disappeared; those remaining in the north are domesticatcd.

Salmon, trout, and *siika* (whitcfish) are rclativcly abundant in the northern rivcrs. Baltic herring is thc most common sea fish, while crayfish can be caught during the brief summer scason. Pikc, char, and perch arc also found.

The vegetation and wildlife of thc Åland Islands arc much like those of southern Finland.

Traditional regions. Thcre arc thrce principal rcgions in Finland: a coastal plain, an interior lake district, and an interior tract of higher land that rises to the fclls (*tunturi*) of Lapland. The coastal plain compriscs a narrow tract in the south, sloping from Salpausselkä to the Gulf of Finland, the southwest plains of the *lääni* (provincc) of Western Finland, and the broad western coastal lowlands of the region of Ostrobothnia (Pohjanmaa) facing the Gulf of Bothnia. The coastal region has thc most extcnsivc strctches of farmland; this region also has the longest continuous settlement and the largest number of urban ccntres. Associated with it are the offshore islands, which are most numerous in thc Turun archipelago off Turku (Åbo) on the southwest coast. Farther to the north in the Gulf of Bothnia, another group of islands lies off Vaasa (Vasa).

The lake district, with its inland archipelagoes, is the heart of Finland. It has been less subject to external influenccs than the coastal region, but since the end of World War II its population has increased, and it has becomc considerably industrialized. The northeast and north are the country's areas of expansion and development where many economic and social interests conflict, including, in the far north, the area of *saamelaisalue,* or Sami territory.

The Åland Islands is a region entirely distinct from Finland, not only because of its geographic separation but also because of its seagirt situation. The islands—whose inhabitants are almost entirely Swedish-speaking—are autono-

The Kokemäen River, with the town of Äetsä in the background, in southwestern Finland.
© Rainer K. Lampinen/Panoramic Images

mous, have their own parliament, and fly their own flag. On the islands farming is a more usual occupation than fishing; there are mixed farms, as in the southwest of Finland, but fruit is also grown. Mariehamn (Maarianhamina) is the capital and only large town.

Settlement patterns. Increased industrialization in Finland has steadily raised the proportion of the population living in urban areas; by the late 20th century about three-fifths of the total population lived in cities and towns. Farms are most commonly located in the meadowland regions of the southwest, where the fertile land is suitable for mixed farming. In the north farmers usually concentrate on small dairy herds and forestry. In Finnish Lapland there is some nomadic life based mainly on the reindeer industry.

Major urban settlements

The major urban settlements are all in the southern third of the country, with a large number of cities and towns concentrated on the coast, either on the Gulf of Finland, as is the capital, Helsinki, or on the Gulf of Bothnia, as are Vaasa and Oulu (Uleåborg). The only town of any size in the north is Rovaniemi, capital of the *lääni* of Lapland. Helsinki is the largest city, with a population about three times that of Tampere (Tammerfors) and of Turku (Åbo), the country's capital until 1812. (C.F.S./I.Su.)

THE PEOPLE

Ethnic and linguistic composition. Excavations undertaken in 1996 have led to a radical reconsideration of how long people have inhabited Finland. Finds in a cave near Kristinestad have led some to suggest that habitation of Finland goes back at least 100,000 years. Ancestors of the Sami apparently were present in Finland by about 7000 BC. As other groups began to enter Finland some 3,000 years later, the proto-Sami probably retreated northward. Archaeological remains suggest that this second wave of settlers came from or had contact with what was to become Russia and also Scandinavia and central Europe. Peoples of Uralic (specifically Finno-Ugric) stock dominated two settlement areas. Those who entered southwestern Finland across the Gulf of Finland were the ancestors of the Tavastlanders, the people of southern and western Finland; those who entered from the southeast were the Karelians. Scandinavian peoples occupied the western coast and archipelagoes and the Åland Islands.

Finland has two national languages, Finnish and Swedish, and is officially bilingual. The Swedish-speaking population, found mainly in the coastal area in the south, southwest, and west and in the Åland Islands (where Swedish is the sole official language), is slowly declining and constitutes roughly 5 percent of the total. Nearly all of the remainder speak Finnish; the language is an important nationalist feature, although it is spoken in strong regional dialects. The Sami-speaking minority in the extreme north numbers some 6,000.

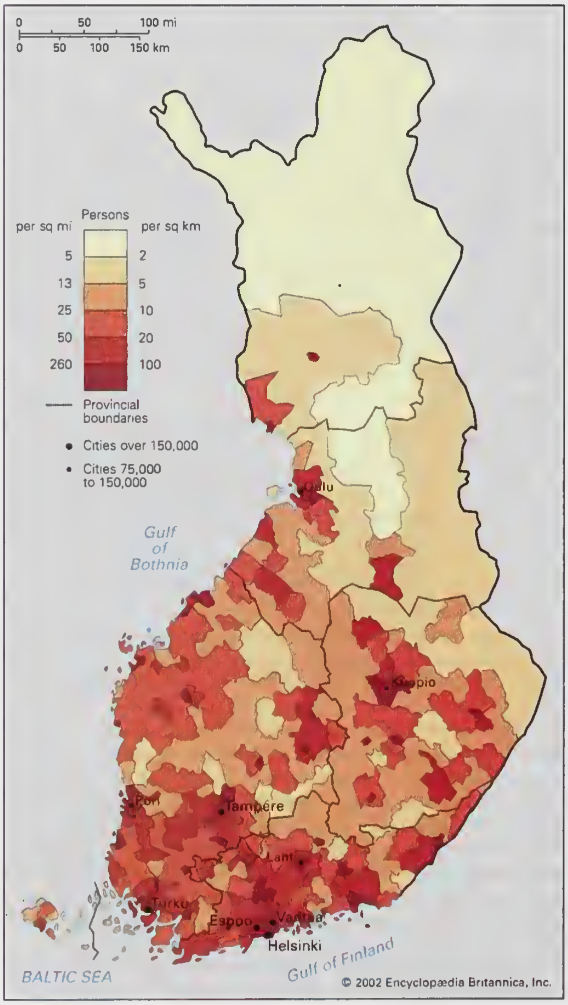
The Finnish Orthodox Church

Religions. Christianity entered Finland from both the west and the east as early as the 12th century. The great majority of the people belong to the Evangelical Lutheran Church, a national church whose bishops are nominated by the head of state. The archbishop has his see at Turku. A small number belong to the Greek Orthodox Church of Finland. The Finnish Orthodox Church was granted autonomy from Moscow in 1920, and in 1923 it was transferred to the jurisdiction of the patriarch of Constantinople. It has one archbishop, with his see at Kuopio. No other Christian denomination in Finland claims more than a few thousand members. Small Jewish and Muslim communities date from about 1850. More than one-tenth of the population have no church affiliation.

Demographic trends. Until the 1990s, emigration exceeded immigration, with Sweden being one of the most attractive destinations for Finnish emigrants. Following World War II, hundreds of thousands of Finns emigrated, while immigration was practically nil, owing to government restrictions. Since 1990, however, Finland has become a country of net immigration. Internal migration since the 1950s has been steadily toward the large towns and cities, most of which are in the south.

THE ECONOMY

Finland's economy is based primarily on private ownership and free enterprise; in some sectors, however, the gov-



Population density of Finland.

ernment exercises a monopoly or a leading role. After World War II, Finland was still only semi-industrialized, with a large part of the population engaged in agriculture, mining, and forestry. During the early postwar decades, primary production gave way to industrial development, which in turn yielded to a service- and information-oriented economy. The economy grew especially rapidly in the 1980s as the country exploited its strong trading relations with both eastern and western Europe. By the early 1990s, however, Finland was experiencing economic recession, reflecting both the collapse of the Soviet Union in 1991 and a general European economic slump. The economy began a slow recovery in the mid-1990s as Finland continued retooling its industry and refocused its trade primarily toward western Europe.

Unemployment was relatively low in Finland until 1991, when it increased rapidly. After peaking at nearly 20 percent of the workforce in 1994, the unemployment rate gradually began to decline again, falling in line with continental trends by the end of the 20th century. Finland's largest employer organization is the Confederation of Finnish Industry and Employers (formerly called the Finnish Employers' Confederation); the largest trade-union groups are the Central Organization of Finnish Trade Unions and the Confederation of Unions for Academic Professionals.

Finland has subscribed to the General Agreement on Tariffs and Trade since 1949 and to the Organisation for Economic Co-operation and Development since 1969. lt became first an associate (1961) and later a full member (1986) of the European Free Trade Association before leaving that organization to join the European Union in 1995; Finland also became a member of the constituent European Community (until 1993 called the European Economic Community), with which it had maintained a free-trade agreement since 1974. The Finnish government derives most of its revenue from income and value-added taxes. About two-fifths of the government's expenditures are for education and social scrvices.

Resources. Trees are Finland's most important natural resource. Some three-fourths of the total land area is forested, with pine, spruce, and birch bcing the prcdominant species. Government cultivation programs, among other measures, have prevented forest depletion; and acid rain, which has devastated forests in central Europe, has not had serious arboreal consequences in Finland.

Finnish peat deposits cover nearly one-third of the country, but only a small fraction of that land is suitable for large-scale peat production. Nearly all the production is used for fuel, with the remainder used in agriculture.

Ore deposits

A diversity of minerals occurs in the Precambrian bedrock, but mining output is modest, owing to the small size of the deposits and the low metal content of the ore. Most mines are located in the north. lron is the most important of the industrial metals. The main nonferrous metals are nickel and zinc. Chromium, cobalt, and copper are also economically important. Gold, silver, cadmium, and titanium are obtained as by-products. There is no naturally occurring coal or oil in Finland. Some mica is quarried, mostly for export.

Agriculture, forestry, and fishing. The declining role of agriculture in Finland's economy is indicated by the steadily decreasing proportion of the labour force working in agriculture. Much land has been taken out of agricultural production, and most farms consist of smallholdings. Finland has been self-supporting in basic foodstuffs since the early 1960s. Meat production roughly equals consumption, while egg and dairy output exceeds domestic needs. Grain production varies considerably; in general, bread grain (mainly wheat) is imported and fodder grain exported. The climate restricts grain farming to the southern and western regions of the country.

Animal husbandry in Finland traditionally concentrated on the raising of dairy cattle, but cuts were made after years of overproduction. As a result, the number of milk cows has declined. The number of horses also declined until the late 1970s but then became generally stable, with the subsequent increase in the number of Thoroughbred horses raised. The keeping of pigs, poultry, and reindeer also is important, while sheep farming and beekecping are of minor economic significance.

Finnish agriculture was heavily subsidized before the country entered the European Union, and after a period of negotiations it remains among the most heavily subsidized under the EU's Common Agricultural Policy. Finnish farmers rely heavily on direct payments based on the amount of land under cultivation. Those north of the 62nd parallel receive especially generous subsidies.

Forest resources

Despite the abundance of forest resources, the forest industry faces increasing production costs. The private owners of more than four-fifths of Finland's forests effectively control domestic timber prices; nonetheless, forest products (notably paper) are a major source of the country's export earnings.

Since World War II, fur farming has made great strides in Finland. Practically all furs are exported; Finland is the world's main producer of farm-raised foxes, and its mink furs also have a high reputation on international markets.

Commercial fishing has gradually become less significant to the economy. Among the fish in Finland's catch are salmon, sea and rainbow trout, whitefish, pike, and char. River pollution, as well as dams built for hydroelectric works, have adversely affected natural spawning habits, especially of salmon and sea trout, and Finland has established a large number of fish-breeding stations at which artificial spawning is induced. There is some trawling for Baltic herring, which also are taken in the winter by seine fishing (dragging nets under the ice) around the offshore islands. There are no longcr any professional hunters in Finland, but recreational hunting remains popular. Elk, waterfowl, and hare are the most common quarries.

Energy. Because of the cold climate and the structure of the nation's industry, Finland's per capita energy consumption ranks among the world's highest. Industries account for about half of total energy consumption, a much higher proportion than the Europcan avcragc. Domcstic energy sources meet only about one-third of Finland's total energy requirement, and all fossil fuels are imported.

Much of Finland's power comes from hydroelectric plants, but the low fall of water makes dam building nccessary. The loss in 1944 of Karelian hydroelectric rcsources turned attention to the north of the country, where plants were built on the Oulu and Kemi rivers. Thermal-generated powcr is also important. Finland's clectricity grids are linked with those of Sweden and Russia, and electricity is imported. lmatran Voima, the state-owned electric power company, began operation of a nuclear plant at Loviisa, east of Helsinki, in 1977; nuclear power now constitutes about one-third of all power generated.

Industry and technology. Finland's northern location imposes certain limitations on industrial activity; severe winter conditions make the costs of construction and heating high, and ice and snow are obstacles to transport. Industrialization in Finland began in the 1860s, but the pacc was slow, and early in the 20th century only some 10 percent of the population derived its livelihood from industry. lt was not until the mid-1960s that industry overtook farming and forestry together as an employer.

Forest products remain a vital sector of the Finnish economy. In the course of development, thc traditional manufactures of vegetable tar and pitch have given way to sawn timber and pulp and later to converted paper products, building materials, and furniture.

Development of heavy industry

Reparations payable to the Soviet Union after World War II, at first a burden, eventually proved a boon to Finland; their payment necessitated the development of heavy industry, which later found markets in western as well as eastern Europe. Metals and engineering now constitute the largest sector of Finnish industry. Finland holds a leading international position in the building of icebreakers, luxury liners, and other specialized ships and in the manufacture of paper-processing equipment. Finland's chemical industry has also grown rapidly. An important branch of the chemical industry is oil refining, the production capacity of which currently exceeds domestic oil requirements.

Textile factories are at Turku, Tampere, Vaasa, Forssa, and Hyvinkää. Helsinki is said to have Europe's largest porcelain factory, while Karhula (Kotka), Iittala, and Nuutajärvi are known internationally for glass. Leather and pewter goods, beer and vodka, and cement are among other important products. Liqueurs, soft drinks, and various sweets are made from cloudberries, currants, gooseberries, and lingonberries.

Finland embraced new technological developments with great enthusiasm. The telecommunications and information technology industry, led by firms such as Nokia, was crucial to the upswing in the Finnish economy of the late 1990s. By 1998 Finland had the largest per capita number of Internet users and mobile telephones in the world. Indeed, "Where are you?" has replaced "Hello" as the standard phone greeting.

Finance. From 1980 the Finnish financial market underwent rapid change. The state's role in the money market declined, and the economy grew increasingly market-oriented. Foreign banks were first allowed to operate in the early 1980s and to open branch offices in 1991.

The Bank of Finland (Suomen Pankki), established in 1811 and guaranteed and supervised by Parliament since 1868, is the country's central bank. The bank controls the circulation of the markka, the national currency established in 1860. Compared with other European countries, Finland has relatively little currency in circulation because Finns are very accustomed to banking electronically. Deposit banks are organized into three groups: commercial, cooperative, and savings. Securities trading is handled by the Helsinki Stock Exchange; foreign investors were first allowed to trade there in the early 1980s.

Trade. Because of Finland's relatively small domestic market, specialized production, and lack of energy sources, foreign trade is vital. The collapse of the Soviet Union and its loss as Finland's chief trading partner shook the Finnish economy. Trade with Russia, while still significant, has been overshadowed by that with the European Union. The chief trading partners are Germany, Sweden, and the United Kingdom. Although the traditional exports of paper and paper products and wood products remain important, heavy machinery and manufactured products now constitute the largest share of Finland's export trade. Imports consist mainly of raw materials for industrial use, consumer goods, and mineral fuels.

Transportation. Until the mid-20th century the problems posed to internal communications and transport by Finland's difficult terrain and weather conditions had hardly been tackled, and many communities remained isolated. External communications were mainly by sea, which, especially as a result of the period of Swedish rule, accounts for the series of well-developed ports on the Gulf of Bothnia and the Gulf of Finland.

Finland has a good network of highways and roads—of which about two-thirds are paved—but the lakes tend to make routes indirect in the southeast, while north of the Arctic Circle the roads are still few. Bridges and car ferries assist road travel in the lakeland areas and in the island archipelagoes. The bus system is highly developed throughout Finland and is widely utilized.

The railway system is much less adequate than that of the roads; the southwestern part of the country is the best-served area. The railways, which provide connections with Russia, are state-owned; about one-third of the rail lines are electrified. In 1982 Finland's first subway was inaugurated in Helsinki.

Navigable waterways

Finland has an extensive network of navigable waterways comprising lakes, rivers, and canals. Many thousands of miles of additional waterways are suitable for the flotage of felled timber, but truck and rail transport is rendering this practice obsolete in many areas. In 1963 the Soviet Union leased to Finland the Soviet end of the canal linking Lake Saimaa with the Gulf of Finland; it was opened in 1968. Most of Finland's overseas cargoes are carried in its own merchant marine. Finland has a passenger liner service, and car ferries operate to Denmark, Sweden, Germany, Estonia, Russia, and Poland.

In addition to the international air terminal near Helsinki, Finland has domestic airports, the most northerly of which is at Ivalo, at Lake Inari. Finnair, the national airline, offers internal and international service.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Finland adopted a republican constitution in 1919; it has been amended several times, notably in the mid-1990s. Legislative power rests in the unicameral Parliament (Eduskunta), which consists of 200 members elected for four years, and in the president, whose term is six years. Executive power is shared by the president and the Council of State, or cabinet, at whose meetings the president takes the chair. The president appoints the prime minister and the cabinet. A clause in the constitution stresses that government ministers are responsible to Parliament.

The role of the president

The president's six-year term of office and the possibility of reelection enhance his powers and provide the country with an important source of stability, in view of the frequent changes of government caused by the multiparty system. In cases of complete deadlock, the president can appoint a nonpolitical caretaker government. Government bills can be introduced into Parliament in the president's name; the president can refuse to sign a bill but must endorse it if it is passed in a subsequent Parliament. The president can dissolve Parliament, has certain decree-making powers, and is the head of the armed forces. The president also conducts foreign policy, but major treaties and matters of war must be validated by Parliament.

Regional and local government. Finland is divided into six *läänit* (provinces)—Åland (Ahvenamaa), Southern Finland (Etelä-Suomi), Eastern Finland (Itä-Suomi), Western Finland (Länsi-Suomi), Lapland (Lappi), and Oulu—each under a governor (*maaherra*) appointed by the president. The provincial governor is in charge of the provincial office (*lääninhallitus*) and the local sheriffs (*nimismies*). Åland, in addition to having a provincial governor, has its own local council, elected by universal suffrage, and a county executive board. The provinces of Finland are divided into communes, which may be rural or urban in character. Each commune council, elected for a four-year term, chooses its executive board. Communes are responsible for local health, education, and social services.

Political parties

Elections, political parties, and trade unions. Suffrage is universal for those over age 18. Presidential elections are direct. There are always more than two presidential candidates. Parliamentary elections are conducted by a system of proportional representation. There are 15 electoral districts. Proportional representation has led to a proliferation of political parties, including the Social Democratic Party, the Left-Wing Alliance (formed in 1990 from the People's Democratic League and the Finnish Communist Party), the National Coalition Party, and the Centre Party (or Finnish Centre; formerly the Agrarian Union). The People's Democratic League and its successor have been important parts of the government since World War II. The Swedish People's Party has become a distinct minority party, though it was formerly more important. The environmentalist Green Union displaced the Finnish Rural Party, a splinter of the former Agrarian Union.

Justice. The judiciary is independent of the legislature and executive; judges are removable only by judicial sentence. There are municipal and rural district courts, held in cities and towns by the chief judge (*oikeuspormestari*) and assistants and in the country by a judge and jurors. Appeal from these courts lies to courts of appeal in Helsinki, Turku, Vaasa, Kuopio, Kouvola, and Rovaniemi. The Supreme Court (Korkein oikeus), in Helsinki, appoints the district judges and those of the appeal courts. The chancellor of justice (*oikeuskansleri*), the supreme judicial authority, also acts as public prosecutor. Parliament appoints a solicitor general, who acts as an ombudsman. The Supreme Administrative Court (Korkein hallinto-oikeus) is the highest tribunal for appeals in administrative cases.

Armed forces. By the Treaty of Paris (1947), Finland was permitted to maintain an army of 34,400 men, an air force of 3,000 men and 60 combat aircraft, and a navy of 4,500 men with ships totaling 10,000 tons. According to the treaty, bombers, submarines, and missiles, as well as nuclear weapons, were forbidden, but in a 1963 agreement Finland was allowed to acquire certain classes of defensive missiles. All male citizens age 17 to 60 are liable for military service, but civil service duty is available to conscientious objectors. (C.F.S./I.Su./S.R.L.)

Education. School attendance in Finland is compulsory beginning at the age of seven. The national and local governments support the schools, and tuition is free. The introduction of a new nine-year comprehensive school system, consisting of a six-year primary stage and a three-year secondary stage, was completed during the 1970s. The English language is taught beginning in the third year, but students can also have the choice of studying other foreign languages. Finland's nine-year comprehensive school system is followed by either a three-year upper secondary school or a vocational school.

University of Helsinki

The only higher education institutions in Finland that were founded before the country achieved independence are the University of Helsinki, founded at Turku in 1640 and transferred to Helsinki in 1828, and the Helsinki University of Technology, founded in 1849. There are also universities located at Jyväskylä, Oulu, Joensuu, Kuopio, and Tampere and two (one Finnish- and one Swedish-language) universities located at Turku. Other institutions of higher education include several technical universities and schools of commerce, as well as a college of veterinary medicine. State aid for higher education is available.

Health and welfare. Social security in Finland comprises a system of pensions and care for the aged, unemployment benefits, health care, and family welfare plans. The state pays disability pensions and old-age pensions to persons 65 years of age and older. The cost of these pensions is met from premiums originally paid by the beneficiaries and payments by employers and by the central and local governments. The Central Pensions Security Institute administers an additional earnings-related old-age pension, which is also available to farmers and other self-employed people. The National Board of Social Welfare provides care and attention for the elderly, including recreational centres to provide social amenities. Other social programs include unemployment benefits and compensation for industrial accidents, maternity benefits, and family allowances for all children under 16.

Health centres, run by local authorities, supply free medical treatment, but there are also licensed private practitioners. The country is divided into hospital districts, each with a central hospital maintained by intermunicipal corporations. There are also smaller regional hospitals and a few private hospitals. The patient pays only a small daily hospital charge. In addition the state reimburses an average of 60 percent of the patient's expenditures on drugs. The Finns are known as a healthy and vigorous people and are characterized by their penchant for sauna baths. Indeed, the life expectancy for Finns is among the highest in the world.

The National Board of Housing addresses problems of housing supply and development. There is a general housing shortage, acute in the towns and especially so in Helsinki. Low-income families are eligible to obtain state-subsidized flats, and government loans for mortgages are also obtainable. Brick and concrete are surpassing wood as building materials, although many Finnish families have vacation cottages, typically modest lakeside dwellings of traditional log or timber construction.

The police authorities are subordinated by the Ministry of the Interior. The cities pay to the state a part of the expenses for local police forces.

CULTURAL LIFE

The *Kalevala*, Finland's national epic

The arts. Finland's national epic, the *Kalevala*, compiled in the 19th century by the scholar Elias Lönnrot from old Finnish ballads, lyrics, and incantations, played a vital part in fostering Finnish national consciousness and pride. Indeed, the development of almost all Finland's cultural institutions and activities has been involved with and motivated by nationalist enthusiasm. This theme can be demonstrated in the growth and development of Finnish theatre and opera, in writing and music, in art and architecture, and also in sport. The festivals of various arts, held annually at places such as Helsinki, Vaasa, and Kaustinen, and the postwar proliferation of museums in Finland show an awareness of the individuality and importance of Finland's contribution to world culture. Savonlinna, in particular, is celebrated for its annual opera festivals.

Theatre, opera, and music. Drama in Finland is truly popular in the sense that vast numbers act in, as well as watch, theatrical productions. Besides the approximately 40 theatre companies in which all the actors are professionals, there are some in which a few professionals or even the producer alone are supplemented by amateur performers. There are amateur theatrical companies in almost every commune.

The country's most important theatre is the National Theatre of Finland, established in 1872 with Kaarlo Bergbom as producer and manager; its granite building in Helsinki was built in 1902. There are also several other municipal theatres. One of the most exciting in the country is the Pyynikki Open Air Theatre of Tampere, the revolving auditorium of which can be moved to face any of the natural sets. There are innumerable institutions connected with the theatre in Finland, including the Central Federation of Finnish Theatrical Organizations. There is a wide repertory of Finnish as well as international plays. The Finnish theatre receives some degree of government assistance.

Finnish opera

The main centre for opera is the Finnish National Opera in Helsinki; the Savonlinna Opera Festival takes place every summer. The international success of Finnish singers such as Taru Valjakka, Jorma Hynninen, and Martti Talvela has added to the continuing national enthusiasm for opera. Several new Finnish operas, including *The Last Temptations* by Joonas Kokkonen and *The Horseman* by Aulis Sallinen, have received successful premieres. Sallinen's *The King Goes Forth to France* (1984) was commissioned jointly by The Royal Opera Covent Garden, the British Broadcasting Corporation, and the Savonlinna Opera Festival.

The dominant figure in Finnish music during the first half of the 20th century was Jean Sibelius, the country's best-known composer, who brought Finnish music into the repertoire of concert halls worldwide. The centre for higher musical studies in Helsinki was renamed the Sibelius Academy. The city is also the location of the Helsinki Philharmonic Orchestra and the Finnish Radio Symphony Orchestra. The Sibelius violin competition and Mirjam Helin song competition are held there every five years. There are annual music festivals in Helsinki and several other cities. Internationally known Finnish conductors include Paavo Berglund, Okko Kamu, and Esa-Pekka Salonen.

Literature. Epic prose has played and continues to play an important role in Finnish literature. *Seitsemän veljestä* (1870; *Seven Brothers*) by Aleksis Kivi is considered to be the first novel written in Finnish. Other early leading prose writers included Frans Eemil Sillanpää, the winner of the Nobel Prize for Literature in 1939. Although Mika Waltari represented newer trends in literature, it was his historical novels, among them *Sinuhe, egyptiläinen* (1945; *The Egyptian*), that brought him fame. Väinö Linna, a leading postwar writer, became known for his war novel *Tuntematon sotilas* (1954; *The Unknown Soldier*) and for the trilogy *Täällä Pohjantähden alla* (1959–62; "Here Under the North Star"). Other novelists have written in shorter forms, but the broad epic has remained popular, particularly among writers describing the contradictions in Finnish life from the turn of the century to modern times.

Literature written in Swedish has had a long tradition in Finland. Among 19th-century writers, Johan Ludvig Runeberg, the national poet, and Zacharias Topelius played leading roles. Later 20th-century poets such as Edith Södergran had a strong influence on the modern poetry of both Finland and Scandinavia. The Swedish language continues to be used in Finnish literature.

Art, architecture, and design. From the time that the *Kalevala* inspired the paintings of Akseli Gallén-Kallela, there has been a distinctive school of Finnish painters, but the Finnish artistic genius has been continually drawn to three-dimensional work. Sculpture is important, highly abstract, and experimental; Eila Hiltanen's monument to Sibelius in Helsinki is composed of chrome, metal, and steel tubes.

Modern Finnish architecture is among the most imaginative and exciting in the world. Its development was closely allied to the nationalist movement, and among its pioneers were Eliel Saarinen, whose work is exemplified by the National Museum and the Helsinki railway station, and Lars Sonck, whose churches in Helsinki and Tampere are particularly notable.

Modern architecture

In the 20th century the idea of functionalism was developed by Gustaf Strengell. In the 1920s Alvar Aalto and Eric Bryggman began experimenting with regional variations on the International Style. Among the most striking examples of Aalto's work are the Paimio Sanatorium, the library at Viipuri, and Finlandia House, a concert and congress hall in Helsinki. There is general experimentation, using concrete and metals, in Finnish industrial buildings and flats and in environmental design, as at the garden town of Tapiola outside Helsinki. The new generation of architects has continued these standards. Outside Finland, Viljo Rewell has received acclaim for his Toronto City Hall in Canada.

Finnish design—especially in glass, porcelain, and textiles—became internationally known during the postwar period. Factories like the well-known Arabia in Helsinki have given artists a free hand to develop their ideas and skills. Tapio Wirkkala and Timo Sarpaneva in glassware, Marjatta Metsovaara in textiles, and Dora Ljung in *ryijy* rugs are among the best-known designers.

Press and broadcasting. Freedom of the press is guaranteed by the 1919 constitution and the Freedom of the Press Act (also 1919); both contain provisions safeguarding editorial rights and outlining press responsibilities. The Supreme Court can suppress publications under certain circumstances, but in general there are few restrictions apart from those governing libel and copyright.

Newspaper publication began in Finland in 1771, and the *Åbo Underrättelser,* published in Swedish, has been in operation since 1824. Finland now has more than 200 regular newspapers, including some 50 dailies, most of which are independently owned and operated. A small number of publications are, however, owned by political parties or trade unions. The national Finnish News Agency (Oy Suomen Tietotoimisto; founded 1887) is independent and owned by the press.

The state-run Finnish Broadcasting Company (Oy Yleisradio Ab [YLE]; established 1926) has no legal monopoly, though no private radio broadcasting companies were allowed until 1985. YLE operates two nationwide television networks and leases a third to a private commercial company; in addition, YLE maintains a fourth channel that broadcasts programs from Swedish television to the coastal areas.

Cross-country skiing

Recreation. In Finland the basic national sport—which originally was a necessary means of winter transportation—is cross-country skiing. Nationalism also encouraged the development of special proficiency, which was fostered by ski fairs and competitions held at Oulu beginning in the late 1890s. An interest in other athletics developed from the time that the Finns took part in the interim Olympic Games held in Athens in 1906. Finns have excelled in Olympic track and field as well as winter sports. Paavo Nurmi, who won six gold medals in Olympic middle- and long-distance running events in the 1920s, became a national hero. Other popular sports are waterskiing, riding, fishing, shooting, ice hockey, and *pesäpallo,* a Finnish version of baseball. (C.F.S./I.Su.)

For statistical data on the land and people of Finland, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

EARLIEST PEOPLES

The first people arrived in Finland about 9,000 years ago. They probably represented several groups and tribes, including the ancestors of the present Sami. Lured by the plenitude of game, particularly fur-bearing animals and fish, they followed the melting ice northward. The first people perhaps came to hunt only for the summer, but gradually more and more of them stayed over the winter. Apparently berries played a significant role in their diet.

Another group probably arrived some 3,000 years later from the southeast. They possibly spoke a Finno-Ugric language and may have been related to the ancestors of the present Finns, if they were not actually of the same group. Other peoples—including the ancestors of the Tavastians—followed from the southwest and central Europe, eventually adopting the Finno-Ugric tongue.

During the 1st millennium BC several more groups arrived, among them the ancestors of the present Finns. The nomadic Sami, who had been scattered over the greater part of Finland, withdrew to the north. Most other groups intermarried and assimilated with the newcomers, and settlement spread across the south of Finland. The population was still extremely sparse, but three loose unities seem to have crystallized: the Finns proper, the Tavastians, and the Karelians. These each had their own chiefs, and they waged war on one another.

Trade colonies

Even before the beginning of the Viking Age (8th–11th century AD), Swedes had settled on the southwestern coast. During the Viking Age, Finland lay along the northern boundary of the trade routes to Russia, and the inhabitants of the area served as suppliers of furs. The Finns apparently did not take part in the Viking expeditions. The end of the Viking Age was a time of unrest in Finland, and Swedish and Danish raids were made on the area, where Russians and Germans also traded.

COMPETITION FOR TRADE AND CONVERTS

From the 12th century, Finland became a battleground between Russia and Sweden. The economic rivalry of the powers in the Baltic was turned into a religious rivalry, and the Swedish expeditions took on the character of crusades. Finland is mentioned together with Estonia in a list of Swedish provinces drawn up for the pope in 1120, apparently as a Swedish missionary area. The first crusade, according to tradition, was undertaken in about 1157 by King Erik, who was accompanied by an English bishop named Henry. Henry remained in Finland to organize the affairs of the church and was murdered by a Finnish yeoman; by the end of the 12th century, he was revered as a saint, and he later became Finland's patron. In a papal bull (*c.* 1172), the Swedes were advised to force the Finns into submission by permanently manning the Finnish fortresses in order to protect the Christianization effort from attacks from the east.

By the end of the 12th century, competition for influence in the Gulf of Finland had intensified: German traders had regular contacts with Novgorod via Gotland, and Denmark tried to establish bases on the gulf. The Danes reportedly invaded Finland in 1191 and again in 1202; in 1209 the pope authorized the archbishop of Lund to appoint a minister stationed in Finland. The Swedish king counterattacked, and in 1216 he received confirmation from the pope of his title to the lands won by himself and his predecessors from the heathens. He was also authorized to establish a seat for one or two bishops in the Finnish missionary territory. In eastern Finland the Russian church attempted to win converts, and in 1227 Duke Jaroslav undertook a program of forced baptisms, designed to tie Karelia closer to Novgorod. In response the pope placed Finland under apostolic protection and invoked a commercial blockade against Russia (1229). A large force, led by Birger, a Swedish jarl (a noble ranking immediately below the king), and including Swedes, Finns, and crusaders from various countries, was defeated in 1240 by a duke of Novgorod, and the advance of Western Christendom into Russia was halted, while the religious division of Finland was sealed, with the Karelians in the Eastern sphere. The bishop of Finland, Thomas, resigned in 1245, and the mission territory was left without leadership until 1249, when the Dominicans founded a monastery in Turku.

FINLAND UNDER SWEDISH RULE

Birger Jarl's conquest

Birger Jarl decided that a full effort was necessary to bring Finland into the Swedish sphere; in 1249 he led an expedition to Tavastia (now Häme), an area already Christianized. Birger built a fortress in Tavastia and some fortifications along the northern coast of the Gulf of Finland, where Swedish settlement on a mass scale began.

Swedes also moved to the eastern coast of the Gulf of Bothnia. In 1293 Torgils Knutsson launched an expedition in an attempt to conquer all of Karelia and built a fortress in Viipuri. The war lasted until 1323, when the Treaty of Pähkinäsaari (Nöteborg; now Petrokrepost) drew the boundary between the Russian and Swedish spheres of influence in a vague line from the eastern part of the Gulf of Finland through the middle of Karelia northwest to the Gulf of Bothnia, and the crusades were ended, with Finland a part of the Swedish realm.

The Swedes began to administer Finland in accordance with Swedish traditions. Castles were built and taxes were collected, mainly in furs and, later, in grain, butter, and money. During the early Middle Ages, Finland was often given to members of the royal family as a duchy. Two new estates, the clergy and the nobility, evolved, with the nobility increased by transplantation from Sweden and the clergy containing a large native element. The first native bishop was appointed in 1291.

Union with Sweden. In 1362 King Haakon of Sweden established the right of the Finns to participate in royal elections and the equal status of Finland with the other parts of the kingdom. Several years later Haakon was overthrown and Albert of Mecklenburg was crowned. Albert was unpopular with the Finns, and by 1374 a Swedish nobleman, Bo Jonsson Grip, had gained title to all of Finland. Grip died in 1386, and Finland soon after became part of the Kalmar Union. (H.En./M.I.He.)

The 15th, 16th, and 17th centuries. Under Swedish sovereignty the Finnish tribes gradually developed a sense of unity, which was encouraged by the bishops of Turku. Study in universities brought Finnish scholars into direct touch with the cultural centres of Europe, and Mikael Agricola (*c.* 1510–57), the creator of the Finnish literary language, brought the Lutheran faith from Germany. As part of medieval Sweden, Finland was drawn into the many wars and domestic battles of the Swedish nobility. In 1581 King John III raised Finland to the level of a grand duchy to irritate his Russian rival, Tsar Ivan IV the Terrible. Dispute over the Swedish crown, combined with quarrels over social conditions, foreign policy, and religion (Roman Catholic versus Lutheran), led to the last peasant revolt in Europe, the so-called Club War, in 1596–97. The hopes of the Finnish peasants were crushed, and, even when Charles IX, whom the peasants had supported, became king (1604–11), the social conditions did not improve. In the course of the administrative reforms of Gustav II Adolf (1611–32), Finland became an integral part of the kingdom, and the educated classes thereafter came increasingly to speak Swedish.

Peace of Täysinä

On its eastern frontier Finland was harassed by constant warfare, and the danger became more serious when Novgorod, at the end of the medieval period, was succeeded by a more powerful neighbour, the Grand Duchy of Moscow. In 1595, however, by the Peace of Täysinä, the existing de facto boundary, up to the Arctic Ocean, was granted official recognition by the Russians. By the Peace of Stolbovo (Stolbova; 1617), Russia ceded Ingermanland and part of Karelia to the kingdom of Sweden-Finland. The population of the ceded territories was of the Greek Orthodox faith, and when the Swedish government began forceful conversion to Lutheranism many fled to Russia and were replaced by Lutheran Finns. After Stolbovo, Sweden found new outlets for expansion in the south and west and developed into one of the leading powers of Europe. Though Finnish conscripts played their part in making Sweden a great power, the role of Finland in the kingdom steadily decreased in importance.

The 18th century. In Charles XII's reign, Sweden lost its position as a great power. During the Great Northern War, Russians occupied Finland for eight years (1713–21), and, under the Peace of Uusikaupunki (Nystad) in 1721, Sweden had to cede the southeastern part of Finland with Viipuri as well as the Baltic provinces. Sweden's capacity to defend Finland had weakened, and the years of hostile occupation had given the Finns a permanent feeling of insecurity.

In the course of the next Russo-Swedish War (1741–43), the Russian empress Elizabeth declared to the Finnish people her intention of making Finland a separate state under Russian suzerainty, but she failed to follow up the idea and at the peace settlement of Turku in 1743 contented herself with annexing a piece of Finland. Meanwhile, however, her original idea had found favour with some Finns. During the next bout of hostilities (1788–90), a number of Finnish officers involved themselves in the activities of Göran Magnus Sprengtporten, a Finnish colonel who had fled to Russia and who wanted to detach Finland from Sweden; this movement won little general support, however.

AUTONOMOUS GRAND DUCHY

As a part of the Swedish monarchy, Finland had been accorded practically no institutions of its own, but from the middle of the 18th century the majority of officials and intellectuals were of Finnish origin. In those circles there was a growing feeling that Finland had to bear the cost of Swedish extravagances in foreign policy. The feeling was not unfounded. Swedish strategic directives of 1785 implied that, in case of Russian attack, Swedish forces should retire from the frontier, leaving Finnish detachments behind, and that under extreme danger the whole of Finland should be evacuated. This strategy was put into effect in 1808–09. Even the treachery of the Anjala association in 1788 was repeated in 1808, when Sveaborg (Viapori; now Suomenlinna) near Helsinki capitulated to the Russians. In 1809 the Finns themselves had to carry the responsibility of coming to terms with Russia. Alexander I offered to recognize constitutional developments in Finland and to give it autonomy as a grand duchy under his throne. (G.Sa./M.I.He.)

The era of bureaucracy. The political framework of Finland under Russia was laid down by the Porvoo (Borgå) Diet in 1809. Finland was still formally a part of Sweden until the peace treaty of Hamina (Fredrikshamn) later that year, but most of the Finnish leaders had already grown tired of Swedish control and wanted to acquire as much self-government as possible under Russian protection. In Porvoo, Finland as a whole was for the first time established as a united political entity—a nation.

In recognition of Finnish autonomy, Alexander I promised to respect the religion and fundamental laws of Finland, as well as the privileges and rights of the inhabitants (that is to say, the Swedish constitution of 1772 as amended in 1789, by which the regent alone had the executive power while the consent of the Diet was required for legislation and the imposition of new taxes). The grand duke (the emperor) was not obliged to convene the Diet at regular intervals, and as a result it did not meet until 1863. From 1809 to 1863 Finland was ruled by a bureaucracy chosen by the Russian emperor, who was represented in Finland by a governor-general. Some holders of this office were Finns in the early period of the Russian regime. The highest administrative organ during the period was the Senate, which consisted of a judicial department and an economic department. The former was the country's supreme court, while the latter became a sort of ministry. A ministerial state secretary in St. Petersburg represented Finnish affairs to the emperor.

Reforms of the Russian period. For most Finns the "era of bureaucracy" was a time of growing prosperity, favourable economic conditions, and no warfare except during the Crimean War (in Finland, the War of Åland). At that time an Anglo-French fleet attacked the Åland Islands, the fortress of Viapori in Helsinki, and some coastal towns on the Gulf of Bothnia. On its separation from Sweden in accordance with the Treaty of Hamina, Finland had a population of more than 900,000. As elsewhere in the Nordic countries, population growth was rapid, and by 1908 the figure had exceeded 2,000,000. Most of the population lived off the land. Manufacture of wooden articles, export of timber, shipbuilding, and merchant shipping were practiced in the small coastal towns.

Despite the strongly authoritarian and bureaucratic form of government, a number of important reforms were implemented. In 1812 the Emperor was induced to restore those areas of Finnish territory that Sweden had ceded to Russia by the treaties of Uusikaupunki (1721) and Turku

(1743). Furthermore, in 1812 Helsinki was chosen as the capital, and the monumental buildings in its centre stem from this period. But the vast rural population and purely agrarian structure prevented the spread of liberal and national ideas to any great extent during the first part of the 19th century.

The Finnish language ordinance

The language problem. The reaction reached its climax with the Finnish language ordinance of 1850, which forbade the publication in Finnish of books other than those that aimed at religious edification or economic benefit. Since Finnish was the only language understood by the majority of the population, the ordinance smacked of an attempt to maintain class differences and was well suited to preserve the existing bureaucracy.

As late as the mid-19th century, Swedish was the only language allowed within the Finnish administration. There was an almost total lack of literature in Finnish, and teaching at both the secondary and university levels was in Swedish. The division between the two languages became not only of national and cultural significance but also a social distinction. This is one of the reasons why the language controversy in Finland created such bitterness. To begin with, the advocates of a Finnish-speaking Finland, or Fennomans, were successful. By recording folk songs and writings, a Finnish literature was developed during the latter part of the 19th century. The first purely Finnish-speaking grammar school appeared in 1858. In 1863 Alexander II (ruled 1855–81) issued a decree stating that, after a 20-year interim period, Finnish was to be placed on an equal footing with Swedish in the administration and in the law courts, as far as their relations with the public were concerned. Swedish, however, remained the language of internal administration, and it was not until 1902 that Swedish and Finnish were placed on an equal footing as official languages.

Reform of the Diet and other reforms. During the reign of Alexander II other reforms were begun. The most important was his convening of the Diet in 1863, and the promulgation of a new act in 1869 providing that it thereafter should be convened regularly. The next great reform period came after the Russian defeat in the war against Japan (1904–05).

Until the 1890s, Russia respected Finland's special position within the Russian Empire in all essentials. In addition to the Diet ordinance of 1869, the country acquired its own monetary system (1865), and a law on conscription, which laid the foundations for the Finnish Army, was passed in 1878.

THE STRUGGLE FOR INDEPENDENCE

Nationalism had already begun to raise its head in Russia before the end of Alexander II's reign, but his strong-minded successor, Alexander III, who had a personal liking for Finland, was able to resist the demands of the Russian nationalists for the abolition of Finnish autonomy and the absorption of the Finns into the Russian nation. The emergence of a united Germany south of the Baltic also worried the Russians, who wanted to secure the loyalty of Finland. Russian jurists took the line that, though Alexander I in virtue of his supreme powers had granted Finland autonomous rights, any Russian emperor exercising the same supreme powers was entitled to take them back whenever he wished. Applying this principle, Nicholas II issued a manifesto on Feb. 15, 1899, according to which he was entitled, without the Finnish Diet's consent, to enact laws enforceable in Finland if such laws affected Russian interests. Direct attempts at Russification were then made. The gradual imposition of Russian as the third official language was ordered in 1900, and in 1901 it was decreed that Finns should serve in Russian units and that Finland's own army should be disbanded. Increasing executive power was conferred on the ultranationalist governor-general, General Nikolay Bobrikov. Faced with this situation, two opposing factions crystallized out of Finland's political parties: the Constitutionalists (the Swedish Party and the Young Finnish Party), who demanded that no one observe the illegal enactments; and the Compliers (the Old Finnish Party), who were ready to give way in everything that did not, in their opinion, affect Finland's vital interest. The Constitutionalists were dismissed from their offices and their leaders were exiled. Young men of Constitutionalist views refused to report for service when called, and at last the Emperor had to give in: the Finnish Army remained disbanded, but no Finns were drafted into the Russian Army. A more extreme group, known as the Activists, was prepared to endorse even acts of violence, and Bobrikov was assassinated by them.

Resistance and reform. Further opposition came from the Labour Party, which was founded in 1899 and which in 1903 adopted Marxist tenets, changing its name to the Social Democratic Party. Unwilling to compromise with tsarist Russia, the party was developing along revolutionary lines. When the Constitutionalists, availing themselves of Russia's momentary weakness, combined with the Social Democrats to organize a national strike, the Emperor restored the situation that had prevailed before 1899 (Nov. 4, 1905)—but not for long. Another result of the strike was a complete reform of the parliamentary system (July 20, 1906). This had been the Social Democrats' most insistent demand. The old four-chamber Diet was changed to a unicameral Parliament elected by equal and universal suffrage. Thus, from having one of Europe's most unrepresentative political systems, Finland had, at one stroke, acquired the most modern. The parliamentary reform polarized the political factions, and the ground was laid for the modern party system. The introduction of universal and equal suffrage meant that the farmers and workers potentially commanded a great majority. The Social Democrats became the largest party in Parliament, obtaining 80 seats out of 200 in the very first elections (1907). Nevertheless, the importance of Parliament remained very small, as it was constantly being dissolved by the Emperor; thus the assault on Finnish autonomy soon began afresh. The Constitutionalists resigned from the government, and the Compliers soon followed their example, since even in their opinion the extreme limit had been overstepped. In the end an illegal Senate composed of Russians was formed. In 1910 the responsibility for all important legislation was transferred to the Russian Duma.

Effect of the Russian Revolution

Return to autonomy. During World War I the Finnish liberation movement sought support from Germany, and a number of young volunteers received military training and formed the Jägar Battalion. After the Russian Revolution in March 1917, Finland obtained its autonomy again, and a Senate, or coalition government, assumed rule of the country. By a law of July 1917 it was decided that all the authority previously wielded by the emperor (apart from defense and foreign policy) should be exercised by the Finnish Parliament. After Russia was taken over by the Bolsheviks in November 1917 Parliament issued a declaration of independence for Finland on Dec. 6, 1917, which was recognized by Lenin and his government on the last day of the year.

EARLY INDEPENDENCE

Although the liberation from Russia occurred peacefully, Finland was unable to avert a violent internal conflict. After the revolutionaries had won control of the Social Democratic Party, they went into action and on Jan. 28, 1918, seized Helsinki and the larger industrial towns in southern Finland. The right-wing government led by the Conservative Pehr Evind Svinhufvud fled to the western part of the country, where a counterattack was organized under the leadership of General Carl Gustaf Mannerheim. At the beginning of April the White Army under his command won the Battle of Tampere. German troops also came to the aid of the White government and captured Helsinki; by May the rebellion had been suppressed. It was followed by trials in which harsh sentences were passed. During the summer and fall of 1918 some 20,000 former revolutionaries either were executed or died in prison camps, bringing the total losses of the war to more than 30,000 lives. A few of the revolutionary leaders, however, managed to escape to Soviet Russia, where a small contingent founded the Finnish Communist Party in Moscow; others continued their flight to the United States and western Europe, some gradually returning to Finland.

Political change. When the Civil War ended, it was

The new constitution

decided, during the summer of 1918, to make Finland a monarchy, and in October the German prince Frederick Charles of Hessen was chosen as king. With Germany's defeat in the war, however, General Mannerheim was designated regent, with the task of submitting a proposal for a new constitution. As it was obvious that Finland was to be a republic, the struggle now concerned presidential power. The liberal parties and the reorganized Social Democratic Party wanted power to be invested in Parliament, while the Conservatives wanted the president to have powers independent of Parliament. The strong position held by the Conservatives after the Civil War enabled them to force through their motion that the president should be chosen by popularly elected representatives, independent of Parliament, and also that he should possess a great deal more authority, especially regarding foreign policy, than at that time was usual for a head of state. After the new constitution had been confirmed on July 17, 1919, the Social Democrats positioned themselves behind the liberal National Progressive Party leader, Kaarlo Juho Ståhlberg, to make him the first president of Finland and to defeat the Conservative candidate Mannerheim, who had not convinced them of his loyalty to republicanism.

Agrarian reform. During the interwar years Finland, to a much greater extent than the rest of the Nordic countries, was an agrarian country. In 1918, 70 percent of the population was employed in agriculture and forestry, and by 1940 the figure was still as high as 57 percent. Paper and wooden articles were Finland's most important export commodities. By the Smallholdings Law of 1918 and by land reform in 1922, which allowed the expropriation of estates of more than 495 acres (200 hectares), an attempt was made to give tenant farmers and landless labourers their own smallholdings. More than 90,000 smallholdings were created, and since then the independent smallholders, who form the majority of the Agrarian Party (now the Centre Party), have been a major factor in Finnish politics.

Political parties. During Ståhlberg's presidency (1919–25), the right-wing parties and the Agrarian Party held power by means of coalitions. The president tried determinedly to minimize the recriminations of the Civil War, and in the course of time he granted amnesty to those who had received long terms of imprisonment. At the same time, the Social Democratic Party was reorganized under the leadership of Väinö Tanner with an exclusively reformist program. When Tanner in 1926 formed a Social Democratic minority government, which granted a general amnesty, the old differences from the Civil War had been almost eliminated. Lauri Kristian Relander, the Agrarian Party's candidate, was elected president in 1925.

Through the first decade of Finnish independence the Social Democratic Party remained the largest party in the Parliament. In the early 1920s the leftist wing of the Social Democrats separated from the party to preach Communism and succeeded in winning 27 seats at the 1922 election. It later changed its name from Socialist Labour Party to Labour Party, but this did not stop the police from arresting all of its parliamentary representatives for treason on the grounds of the party's revolutionary intent. The Communists, however, once more reorganized and worked closely with the Finnish Communist Party in the Soviet Union. In the following elections they were able to win about 20 seats in Parliament.

The Lapua Movement

As a reaction to the growing Finnish Communist Party, the Lapua (Lappo) Movement emerged and in the years 1929–32 attempted to force its demands through actions against Communist newspapers, acts of terrorism against individual citizens, and mass demonstrations. These actions, which were supported by the Conservatives and many members of the Agrarian Party, were at first successful. The Communists were prevented from taking part in the 1930 election, and the 66 Social Democrats were one too few in the Parliament to prevent the passage of an anti-Communist law. This law banned the public activities of the Communist Party, forced its members underground, and stripped them of their right to vote, virtually eliminating their influence on Finnish politics. In 1931 Svinhufvud was elected president with the help of the Lapua Movement. When the Lapua Movement shortly afterward turned its activities against the Social Democrats, too, and tried to seize power by force in the Mäntsälä coup attempt in 1932, the president intervened and managed in a radio speech to calm the rebellion. Another failure at this time was the law on the total prohibition of alcohol, introduced in 1919. As in the United States, the law resulted in a sharp increase in organized crime and smuggling, and after a referendum in 1932 it was repealed.

The language question. The 1919 constitution provided that both Finnish and Swedish should be the national languages. A younger radical generation now raised the demand for the supremacy of Finnish. The language controversy, which during the interwar period was a very bitterly fought issue, caused the position of the Swedish language to be progressively weakened toward the end of the 1930s. During World War II the laws governing language were revised, first in 1947 and again in 1961. They now guarantee equal status for Swedish, which remains an official language of the country and a required subject in Finnish schools.

Foreign policy. After the recognition of Finland as a sovereign state, two problems had to be faced. The first was in connection with the eastern boundary, where influential groups wished to annex East Karelia. By the Treaty of Tartu (Dorpat) in 1920, however, the boundary was unchanged except in the north, where Finland acquired a route to the Arctic Ocean and the harbour of Petsamo. The other problem concerned the Åland Islands (Finnish: Ahvenanmaa), which Sweden had temporarily occupied during the Finnish Civil War. The demands of the population of the islands to be united with Sweden were firmly rejected. The League of Nations settled the question in 1921 in accordance with Finland's wishes.

Finland's main security problems resulted from the threat from the Soviet Union. An attempt to solve this by a defense alliance with Estonia, Latvia, and Poland in 1922 failed when Parliament refused to ratify the agreement, and in 1932 a Finnish–Soviet nonaggression pact was signed. Despite this, relations between the two countries did not really improve, and they remained "neighbours against their will." During the second half of the 1930s a Finnish–Swedish defense association was planned that, among other things, would have brought about the rearming of Åland, but the Soviet Union objected to these plans and they could not be realized.

FINLAND DURING WORLD WAR II

The Treaty of Moscow

The Winter War. After Poland's defeat in the autumn of 1939, the Soviet Union, wishing to safeguard Leningrad, demanded from Finland a minor part of the Karelian Isthmus, a naval base at Hanko (Hangö), and some islands in the Gulf of Finland. When Finland rejected the demand, the Soviet Union launched an attack on Nov. 30, 1939. Immediately after the attack a coalition government was formed under Risto Ryti. Despite courageous resistance and a number of successful defense actions, the defense of the Karelian Isthmus broke down, and Finland had to initiate peace negotiations. By the Treaty of Moscow of March 12, 1940, Finland surrendered a large area of southeastern Finland, including the city of Viipuri (renamed Vyborg), and leased the peninsula of Hanko to the Soviet Union for 30 years.

Cooperation with Germany. After the Treaty of Moscow the plan for a Nordic defense union was resumed. The Soviet Union still objected, however, and the plan was thus abandoned. In December 1940 President Kyösti Kallio resigned, and Risto Ryti was elected in his place. When the tension between Germany and the Soviet Union grew in spring 1941, Finland approached Germany but did not conclude a formal agreement. Nevertheless, Finland, like Sweden after Norway's capitulation, allowed the transit of German troops. When Germany attacked the Soviet Union on June 22, 1941, therefore, German troops were already on Finnish territory, and Finland was ready for war; its submarines, in fact, were operating in Soviet waters. The "War of Continuation" (1941–44) began with a successful Finnish offensive that led to the capture of large areas of East Karelia. Some Finns were reluctant, however, to cross the old border of 1939, and the spirit of the Winter War that had united the Finns began to weaken. From the winter of 1942–43, Germany's defeats gave rise to a growing demand for peace in Finland. After the breakthrough of the Red Army on the Karelian Isthmus in June 1944, President Ryti resigned on August 1. He was succeeded by Marshal Gustaf Mannerheim, who began negotiations for an armistice. This was signed on Sept. 19, 1944, on condition that Finland recognize the Treaty of Moscow of 1940 and that all foreign (German) forces be evacuated. A pledge was given, moreover, to cede Petsamo; to lease an area near Porkkala, southwest of Helsinki, for a period of 50 years (in place of Hanko); and within 6 years to pay the equivalent of $300 million in goods for war reparations. In the meantime, however, the German army refused to leave the country, and, in a series of clashes that followed, it devastated great areas of northern Finland in its retreat. The final peace treaty, signed in Paris on Feb. 10, 1947, reiterated the conditions of the armistice agreement. (Jö.We./M.I.He.)

THE POSTWAR PERIOD

After the armistice in 1944 a coalition government was formed under the leadership of Juho Kusti Paasikivi. When conditions had been stabilized, Mannerheim resigned, and Paasikivi was elected president in his place in 1946. In 1956 the leader of the Agrarian Party, Urho Kekkonen, who acted as prime minister a number of times during the period 1950 to 1956, was elected president. He was reelected three times to the office, with an extension of his third term by the Parliament. When he resigned in 1981, he was succeeded by the Social Democrat Mauno Koivisto, who was reelected in 1988.

Paasikivi's tenure

Foreign policy. Under the leadership of Paasikivi and Kekkonen, relations with the Soviet Union were stabilized by a consistently friendly policy on the part of Finland. A concrete expression of the new foreign policy was the Agreement of Friendship, Cooperation, and Mutual Assistance concluded between Finland and the Soviet Union in 1948 and extended in 1955, 1970, and 1983. The agreement included a mutual defense provision and prohibited Finland from joining any organization considered hostile to the U.S.S.R. After war reparations had been paid in full, trade with the Soviet Union continued, rising to more than 25 percent of Finland's total during the 1980s. Further signs of the détente showed when the Soviet Union returned its base at Porkkala in 1955.

Relations with the Soviet Union, however, were not entirely without complications. After the elections of 1958, a coalition government under the leadership of the Social Democrat Karl August Fagerholm included certain members considered anti-Soviet. The Soviet Union responded by recalling its ambassador and canceling credits and orders in Finland. When the Finnish government was reconstructed, relations were again stabilized. During the autumn of 1961, when international relations were severely strained because of the Berlin crisis, the Soviet Union requested consultations in accordance with the 1948 agreement. President Kekkonen succeeded in inducing the Soviet Union to abandon its request. In 1985 the Soviets warned that a split in the Finnish Communist Party between the nationalist-reformist majority and the pro-Moscow minority would jeopardize Soviet-Finnish relations, but the split occurred in 1986 without straining relations.

Nordic cooperation. Finland became a member of the United Nations and of the Nordic Council in 1955. Nordic cooperation has led to many shared policies in Finland, Denmark, Iceland, Norway, and Sweden. These include free movement across the borders of these five countries, the gradual development of a common and free labour market, and various political, economic, and cultural measures. In 1986 Finland became a full member of the European Free Trade Association (EFTA).

Domestic affairs. During the early postwar years, Finland confronted economic difficulties. After World War II the country had to absorb about 300,000 refugees from the areas ceded to the Soviet Union and at the same time pay war reparations. Despite these obstacles Finland quickly recovered. The war reparations brought about rapid expansion in the metal and shipbuilding industries, and the timber trade soon resumed exporting and quickly exceeded its prewar level. The rebuilding and colonization required to resettle the refugees, however, were such a strain on the country's economy that inflation ensued; as a result Finland devalued its currency repeatedly.

After the armistice, the new Finnish Communist Party held a strong position. When in the spring of 1948 it was alleged that the party had planned a coup, Parliament forced the Communist minister of the interior to resign. Parliamentary elections in the autumn of 1948 resulted in a Social Democratic government under the leadership of Fagerholm. Governments changed rapidly and consisted of various party coalitions during the 1950s, in most cases under the leadership of the Agrarian Party or the Social Democrats. During this period, however, both the Conservative National Coalition Party and the leftist Finnish People's Democratic League, which included the Finnish Communist Party, were excluded from the government.

Forming and keeping a government in Finland was very difficult because of the proliferation of political parties; no one party, and often no party group, could command a majority in Parliament. As a consequence, there were many nonpolitical cabinets composed of civil servants appointed by the president. With continuing economic growth and because of internal disputes, Communist Party influence diminished after the 1970s, and after the party's split in the mid-1980s the Communists suffered severe losses in the 1987 election. The Conservatives gained and formed a coalition government under Conservative Prime Minister Harri Holkeri.

RECENT DEVELOPMENTS

In 1991 a new cabinet, formed by major nonsocialist parties and with the Centre's Esko Aho as prime minister, immediately faced Finland's worst economic recession since the 1940s. During the early 1990s production dropped sharply and unemployment skyrocketed, largely because trade with Russia plummeted from the levels of the Soviet era. Following the Soviet Union's demise in 1991, Finland moved to end the old mutual defense agreement. A new accord was reached with Russia in 1992, in which the two countries simply pledged to settle disputes between them peacefully. Finland, now freed from any restrictions, applied for membership in the European Community, which formed the European Union (EU) in 1993. When Finland joined the EU in 1995, it left EFTA. Economic recovery came slowly during the mid-1990s, as export markets gradually shifted toward the EU countries. The government tried to cut expenditures, notably on social programs. The public expressed its displeasure with the slow pace of recovery by ousting the Centre from the government in the elections of 1995.

Social Democrat Paavo Lipponen formed a cabinet from a broad-based coalition that included, for the first time, members of the environmentalist Green Union. Lipponen's government moved cautiously toward closer military union with western Europe and a more assertive stance toward Russia. Finland joined other EU countries in adopting the common European currency, the euro, in 1999, even as other Scandinavian countries resisted incorporation into the euro zone. Meanwhile, led by a strong high-technology sector, Finland's economy prospered at the turn of the 21st century. When President Martti Ahtisaari (who had succeeded Koivisto in 1994) left office in 2000, Social Democrat Tarja Halonen became Finland's first woman president. When Russian president Vladimir Putin refused to discuss the return of Karelia, Halonen responded firmly that the matter was not closed. Free at last from fear of Soviet domination, an economically robust Finland enjoyed full independence of action for perhaps the first time in its history. (Jö.We./M.I.He./Ed.)

For later developments in the history of Finland, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 923, 961, 963, and 972.

BIBLIOGRAPHY

General works. Overviews are provided in RIITTA DA COSTA and PAUL KOJO (eds.), *Facts About Finland,* 2nd ed. (1985); SYLVIE NICKELS, HILLAR KALLAS, and PHILIPPA FRIEDMAN (eds.), *Finland: An Introduction,* 2nd rev. ed. (1973); MAX ENGMAN and DAVID KIRBY (eds.), *Finland: People, Nation, State* (1989); ERIC SOLSTEN and SANDRA W. MEDITZ (eds.), *Finland: A Country Study,* 2nd ed. (1990); and *Finland Handbook* (annual), published by the Finnish Tourist Board.

Physical and human geography. W.R. MEAD, *An Historical Geography of Scandinavia* (1981); and KALEVI RIKKINEN, *A Geography of Finland,* trans. from Finnish (1992), provide comprehensive surveys. A broad interpretive treatment, with a look at the social customs of Finland, is found in PHILIP WARD, *Finnish Cities: Travels in Helsinki, Turku, Tampere, and Lapland* (1987).

Ethnological studies include AURÉLIEN SAUVAGEOT, *Les Anciens Finnois* (1961); and WILLIAM A. WILSON, *Folklore and Nationalism in Modern Finland* (1976). Social life and customs are explored in AINI RAJANEN, *Of Finnish Ways* (1981); CAJ BREMER and ANTERO RAEVUORI, *The World of the Sauna* (1986; originally published in Finnish, 1985); ANTTI TUURI, *The Face of Finland,* ed. by PAULI KOJO, trans. from Finnish (1983); and ANNEKE LIPSANEN, *The Finnish Folk Year: A Perpetual Diary & Book of Days, Ways, and Customs* (1987).

Finland's economy is discussed in FRED SINGLETON, *The Economy of Finland in the Twentieth Century* (1986); RIITTA HJERPPE, *The Finnish Economy, 1860–1985: Growth and Structural Change* (1989; originally published in Finnish, 1988); ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Reviews of National Science and Technology Policy: Finland* (1987); *Environmental High-Technology from Finland* (1986), published by the Ministry of the Environment; *Economic Survey* (annual), published by the Ministry of Finance; and *Finnish Industry,* rev. ed. (1982), an overview of developments, published by the Bank of Finland.

Government and politics are analyzed in D.G. KIRBY, *Finland in the Twentieth Century* (1979); ANTHONY F. UPTON, PETER P. ROHOE, and A. SPARRING, *Communism in Scandinavia and Finland* (also published as *The Communist Parties of Scandinavia and Finland,* 1973); JUHANI MYLLY and R. MICHAEL BERRY (eds.), *Political Parties in Finland* (1984); JAAKKO NOUSIAINEN, *The Finnish Political System* (1971; originally published in Finnish, 3rd ed., 1967); DAVID ARTER, *Politics and Policy-Making in Finland* (1987); and RISTO ALAPURO, *State and Revolution in Finland* (1988).

Finnish architecture and design are discussed in J.M. RICHARDS, *800 Years of Finnish Architecture* (1978); ERIK KRUSKOPF, *Finnish Design, 1875–1975: 100 Years of Finnish Industrial Design* (1975); ELIZABETH GAYNOR, *Finland, Living Design* (1984, reissued 1995); and JAAKKO LINTINEN *et al., Finnish Vision: Modern Art, Architecture, and Design,* trans. from Finnish (1983). Other studies of national art and culture include JOHN BOULTON SMITH, *The Golden Age of Finnish Art: Art Nouveau and the National Spirit,* 2nd rev. ed. (1985); MARIANNE AAV and KAJ KALIN, *Form Finland,* trans. from Finnish (1986), on decorative arts; JAAKKO AHOKAS, *A History of Finnish Literature* (1973); MATTI KUUSI, KEITH BOSLEY, and MICHAEL BRANCH (eds. and trans.), *Finnish Folk Poetry: Epic: An Anthology in Finnish and English* (1977); KAI LAITINEN, *Literature of Finland: An Outline*, 2nd ed., trans. from Finnish (1994); *Kalevala,* ed. by AIVI GALLEN-KALLELA and trans. by W.F. KIRBY (1986), a jubilee edition of the national epic, illustrated by AKSELI GALLEN-KALLELA; *The Kalevala: Epic of the Finnish People*, trans. by EINO FRIBERG and ed. by GEORGE C. SCHOOLFIELD (1988); ANTONY HODGSON, *Scandinavian Music: Finland & Sweden* (1984); PAAVO HELISTÖ, *Music in Finland* (1980); and MAIJA SAVUTIE, *Finnish Theatre: A Northern Part of World Theatre,* trans. from Finnish (1980).

History. General works on Finnish history include JOHN H. WUORINEN, *A History of Finland* (1965); EINO JUTIKKALA and KAUKO PIRINEN, *A History of Finland,* 4th rev. ed. (1984; originally published in Finnish, 1966); EINO JUTIKKALA, *Atlas of Finnish History,* 2nd rev. ed. (1959); BYRON J. NORDSTROM (ed.), *Dictionary of Scandinavian History* (1986); FRED SINGLETON, *A Short History of Finland* (1989); and MATTI KLINGE, *A Brief History of Finland,* trans. from Finnish, 10th ed. (1994).

More detailed discussions of events in the 19th and 20th centuries are available in JUHANI PAASIVIRTA, *Finland and Europe: International Crises in the Period of Autonomy, 1808–1914,* ed. and abridged by D.G. KIRBY (1981; originally published in Finnish, 1978); L.A. PUNTILA, *The Political History of Finland, 1809–1966* (1974; originally published in Finnish, 5th rev. and improved ed., 1971); ANTHONY F. UPTON, *The Finnish Revolution, 1917–1918* (1980), a comprehensive analysis, and *Finland, 1939–1940* (1974); and MAX JAKOBSON, *Finland Survived: An Account of the Finnish-Soviet Winter War, 1939–1940,* 2nd enlarged ed. (1984).

Foreign relations are the main topic of TUOMO POLVINEN, *Between East and West: Finland in International Politics, 1944–1947,* ed. and trans. by D.G. KIRBY and PETER HERRING (1986; originally published in Finnish, 3 vol., 1979–81); ROY ALLISON, *Finland's Relations with the Soviet Union, 1944–1984* (1985); R. MICHAEL BERRY, *American Foreign Policy and the Finnish Exception* (1987); and MAX JAKOBSON, *Finland: Myth and Reality* (1987). The *Yearbook of Finnish Foreign Policy,* published by the Finnish Institute of International Affairs, is another helpful source. (I.Su./M.I.He./Ed.)

Fishes

The term fish is applied to a variety of cold-blooded aquatic vertebrates of several evolutionary lines. It describes a life-form rather than a taxonomic group. As members of the phylum Chordata, fish share certain features with other vertebrates. These features are gill slits at some point in the life cycle, a notochord, or skeletal supporting rod, a dorsal hollow nerve cord, and a tail. Living fishes represent about five classes, which are as distinct from one another as are the four classes of familiar air-breathing animals—amphibians, reptiles, birds, and mammals. For example, the jawless fishes (agnathans) are the only fishes that have a suctorial, or filter-feeding, mouth, a feature which makes them dependent on an essentially parasitic way of life. They have either no fins or poorly developed ones. Extant examples of the agnathans are the lampreys and the hagfishes. As the name implies, the skeletons of fishes of the class Chondrichthyes (*chondr,* "cartilage," and *ichthyes,* "fish") are made entirely of cartilage. Modern fish of this class lack a swim bladder, and their scales and teeth are made up of the same placoid material. Sharks, skates, and rays are examples of cartilaginous fishes. The bony fishes are by far the largest class. Examples range from the tiny sea horse to the 450-kilogram (1,000-pound) blue marlin, from the flat soles and flounders to the boxy puffers and sunfishes. Unlike those of the cartilaginous fishes, the scales of bony fishes, when present, grow throughout life and are made up of thin, overlapping plates of bone. Bony fishes also have an operculum that covers the gill slits.

The study of fishes, the science of ichthyology, is of broad importance. There are many reasons why fishes are of interest to humans; the most important is their relationship with and dependence on the environment. A more obvious reason for interest in fishes is their role as a moderate but important part of the world's food supply. This resource, once thought unlimited, is now realized to be finite and in delicate balance with the biological, chemical, and physical factors of the aquatic environment. Overfishing, pollution, and alteration of the environment are the chief enemies of proper fisheries management, both in fresh waters and in the ocean. (For a detailed discussion of the technology and economics of fisheries, see FISHING, COMMERCIAL.)

Another practical reason for studying fishes is their use in disease control. As predators on mosquito larvae, they help curb malaria and other mosquito-borne diseases. Fishes are valuable laboratory animals in many aspects of medical and biological research. For example, the readiness of many fishes to acclimate to captivity has allowed biologists to study behaviour, physiology, and even ecology under relatively natural conditions. Fishes have been especially important in the study of animal behaviour where research on fishes has provided a broad base for the understanding of the more flexible behaviour of the higher vertebrates.

There are aesthetic and recreational reasons for an interest in fishes. Millions of people keep live fishes in home aquariums for the simple pleasure of observing the beauty and behaviour of animals otherwise unfamiliar to them. To many, aquarium fishes provide a personal challenge, allowing them to test their ability to keep a small section of the natural environment in their homes. Sportfishing is another way of enjoying the natural environment, also indulged in by millions of people every year. Interest in aquarium fishes and sportfishing support multimillion-dollar industries throughout the world.

As mentioned above, the fishes represent several classes of vertebrates rather than a single taxonomic group. This article presents a comparative study of all fish groups, including those that are now extinct, with emphasis on their structural diversity, natural history, and evolutionary relationships. The major classes of living fishes—arranged in accordance with the complete *Annotated classification* (see outline)—are treated individually. Special attention is devoted to the teleosts (infraclass Teleostei), or bony fishes, a group that includes most of the world's important sport and commercial fishes. (S.H.W./Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 312, and the *Index.*

The article is divided into the following sections:

FISHES: A COMPARATIVE STUDY

GENERAL FEATURES

Structural diversity. Fishes have been in existence for more than 450,000,000 years, during which time they have evolved repeatedly to fit into almost every conceivable type of aquatic habitat. In a sense, land vertebrates are simply highly modified fishes, for when fishes colonized the land habitat they became tetrapod (four-legged) land vertebrates. The popular conception of a fish as a slippery, streamlined aquatic animal that possesses fins and breathes by gills applies to many fishes, but far more fishes deviate from that conception than conform to it. For example, the body is elongate in many forms and greatly shortened in others; the body is flattened in some (principally in bottom-dwelling fishes) and laterally compressed in many others; the fins may be elaborately extended, forming intricate shapes, or they may be reduced or even lost; and the positions of the mouth, eyes, nostrils, and gill openings vary widely. Air breathers have appeared in several evolutionary lines.

Many fishes are cryptically coloured and shaped, closely matching their respective environments; others are among the most brilliantly coloured of all organisms, with a wide range of hues, often of striking intensity, on a single individual. The brilliance of pigments may be enhanced by the surface structure of the fish, so that it almost seems to glow. A number of unrelated fishes have actual light-producing organs. Many fishes are able to alter their coloration, some for the purpose of camouflage, others for the enhancement of behavioral signals.

Fishes range in adult length from less than 10 millimetres (2/5 inches) to more than 20 metres (60 feet) and in weight from about 1.5 grams (less than 1/16 ounce) to many thousands of kilograms. Some live in shallow thermal springs at temperatures slightly above 42° C (100° F), others in cold Arctic seas a few degrees below 0° C (32° F) or in cold deep waters more than 10,000 metres (3,500 feet) beneath the ocean surface. The structural and, especially, the physiological adaptations for life at such extremes are relatively poorly known and provide the scientifically curious with great incentive for study.

Distribution and abundance. Almost all natural bodies of water bear fish life, the exceptions being very hot thermal ponds and extremely salt-alkaline lakes such as the Dead Sea and Great Salt Lake in Utah. The present distribution of fishes is a result of the geological history and development of the Earth as well as the ability of fishes to undergo evolutionary change and to adapt to the available habitats. Fishes may be seen to be distributed according to habitat and according to geographical area. Major habitat differences are marine and fresh waters. For the most part the fishes in them, even in adjacent areas, are different, but some, such as the salmon, migrate from one to the other. The freshwater habitat may be seen to be of many kinds. Fishes found in mountain torrents, Arctic lakes, tropical lakes, temperate streams, and tropical rivers will all differ from each other both in obvious gross structure and in physiological attributes. Even in closely adjacent habitats where, for example, a tropical mountain torrent enters a lowland stream, the fish fauna will differ. Marine habitats can be divided into deep ocean floors (benthic), midwater oceanic (bathypelagic), surface oceanic (pelagic), rocky coast, sandy coast, muddy shores, bays, estuaries, and others. Also, for example, rocky coastal shores in tropical and temperate regions will have a different fish fauna, even when such habitats occur along the same coastline.

Habitats

Although much is known about the present geographical distribution of fishes, far less is known about how that distribution came about. Many parts of the fish fauna of the fresh waters of North America and Eurasia are related and undoubtedly have a common origin. The faunas of Africa and South America are related, extremely old, and probably an expression of the drifting apart of the two continents. The fauna of southern Asia is related to that of central Asia and some of it appears to have entered Africa. The extremely large shore fish faunas of the Indian and tropical Pacific oceans comprise a related complex, but the tropical shore fauna of the Atlantic, although containing Indo-Pacific components, is relatively limited and probably younger. The Arctic and Antarctic marine faunas are quite different from each other. The shore fauna of the North Pacific is quite distinct, and that of the North Atlantic more limited and probably younger. Pelagic oceanic fishes, especially those in deep waters, are similar the world over, showing little geographical isolation in terms of family groups. The deep oceanic habitat is very much the same throughout the world, but species differences do exist, showing geographical areas determined by oceanic currents.

NATURAL HISTORY

Life history. All aspects of the life of a fish are closely correlated with adaptation to the total environment, physical, chemical, and biological. In studies of fish life, all the interdependent aspects of their life, such as behaviour, locomotion, reproduction, and physical and physiological characteristics, must be taken into account.

Correlated with their adaptation to an extremely wide variety of habitats is the extremely wide variety of life cycles that fishes display. The great majority hatch from relatively small eggs a few days to several weeks or more after the eggs are scattered in the water. Newly hatched young are still partially undeveloped and are called larvae until body structures such as fins, skeleton, and some organs are fully formed. Larval life is often very short, usually less than a few weeks, but it can be very long, some lampreys continuing as larvae for at least five years. Young and larval fishes, before reaching sexual maturity, must grow considerably, and their small size and other factors often dictate that they live in a habitat different than that of the adults. For example, some tropical marine shore fishes have pelagic larvae. Larval food also is different and they often live in shallow waters, where they may be less exposed to predators.

Longevity

After the fish reaches adult size, the length of its life is subject to many factors, such as innate rates of aging, predation pressure, and the nature of the local climate. The longevity of a species in the protected environment of an aquarium may have nothing to do with how long members of that species live in the wild. Many small fishes live only one to three years at the most. In a few large species some individuals may live as long as 10 or 20 years or even longer.

Behaviour. Fish behaviour is a complicated and varied subject. As in almost all animals with a central nervous system, the nature of a response of an individual fish to stimuli from its environment depends upon the inherited characteristics of its nervous system, on what it has learned from past experience, and on the nature of the stimuli. Compared with the variety of human responses, however, that of a fish is stereotyped, not subject to much modification by "thought" or learning, and investigators must guard against anthropomorphic interpretations of fish behaviour.

Fishes perceive the world around them by the usual senses of sight, smell, hearing, touch, and taste and by special lateral-line water-current detectors. In the few fishes that generate electric fields, a process that might best be called electrolocation aids in perception. One or another of these senses often is emphasized at the expense of others depending upon the fish's other adaptations. In fishes with large eyes the sense of smell may be reduced; others, with small eyes, hunt and feed primarily by smell (*e.g.,* some eels).

Specialized behaviour is primarily concerned with the three most important activities in the fish's life: feeding, reproduction, and escape from enemies. Schooling behaviour of sardines on the high seas, for instance, is largely a protective device to avoid enemies, but it is also associated with and modified by their breeding and feeding requirements. Predatory fishes are most often solitary, lying in wait to dart suddenly after their prey, a kind of locomotion impossible for beaked parrot fishes, which feed on coral, swimming in small groups from one coral head to the next.

Sleep

Sleep in fishes, all of which lack true eyelids, consists of a seemingly listless state in which the fish maintains its balance but moves slowly. If attacked or disturbed, most can dart away. A few kinds of fishes lie on the bottom to sleep. Most catfishes, some loaches, and some eels and electric fishes are strictly nocturnal, being active and hunting for food during the night and retiring during the day to holes, thick vegetation, or other protective parts of the environment.

Communication between members of a species or between members of two or more species often is extremely important, especially in breeding behaviour (see below *Reproduction*). The mode of communication may be visual, as between the small so-called cleaner fish and a large fish of a very different species. The larger fish often allows the cleaner to enter its mouth to remove gill parasites. The cleaner is recognized by its distinctive colour and actions and therefore is not eaten, even if the larger fish is normally a predator.

Locomotion. Many fishes have a streamlined body and swim freely in the open water. Fish locomotion is closely correlated with habitat and ecological niche (the general position of the animal to its environment).

Many fishes in both marine and fresh waters swim at the surface and have mouths adapted to feed best (and sometimes only) at the surface. Often such fishes are long and slender, able to dart at surface insects or at other surface fishes and in turn to dart away from predators; needlefishes, halfbeaks, and topminnows are good examples. Oceanic flying fishes escape their predators by gathering speed above the water surface, with the lower lobe of the tail providing thrust in the water. They then glide hundreds of yards on enlarged, winglike pectoral and pelvic fins. South American freshwater flying fishes escape their enemies by jumping and propelling their strongly keeled bodies out of the water with their pectoral fins, which function as flapping wings.

So-called midwater swimmers, the most common type of fish, are of many kinds and live in many habitats. The powerful fusiform tunas and the trouts, for example, are adapted for strong, fast swimming, the first to capture prey speedily in the open ocean, the second to cope with the swift currents of streams and rivers. The trout body form is well adapted to many habitats. Fishes that live in relatively quiet waters such as bays or lake shores or slow rivers usually are not strong, fast swimmers but are capable of short, quick bursts of speed to escape a predator. Many of these fishes have their sides flattened, examples being the sunfish and the freshwater angelfish of aquarists. Fish associated with the bottom or substrate usually are slow swimmers. Open-water plankton-feeding fishes almost always remain fusiform and capable of rapid, strong movement (for example, sardines and herrings of the open ocean and also many small minnows of streams and lakes).

Special adaptations of bottom dwellers

Bottom-living fishes are of many kinds and have undergone many types of modification of their body shape and swimming habits. Rays, which evolved from strong swimming, midwater sharks, usually stay close to the bottom and move by undulating their large pectoral fins. Flounders live in a similar habitat and move over the bottom by undulating the entire body. Many bottom fishes dart from place to place, resting on the bottom between movements, a motion common in gobies. One goby relative, the mudskipper, has taken to living at the edge of pools along the shore of muddy mangrove swamps. It escapes its enemies by flipping rapidly over the mud, out of the water. Some catfishes, synbranchid eels, the so-called climbing perch, and a few other fishes venture out over damp ground to find more promising waters than those that they left. They move by wriggling their bodies, sometimes using strong pectoral fins; most have accessory air-breathing organs. Many bottom-dwelling fishes live in mud holes or rocky crevices. Marine eels and gobies commonly are found in such habitats and for the most part venture far beyond their cavelike homes. Some bottom dwellers, such as the clingfishes (Gobiesocidae), have developed powerful adhesive disks that enable them to remain in place on the substrate in areas such as rocky coasts where the action of the waves is great.

Reproduction. The methods of reproduction in fishes are varied, but most fishes lay a large number of small eggs, fertilized and scattered outside of the body. The eggs of pelagic fishes usually remain suspended in the open water. Many shore and freshwater fishes lay eggs on the bottom or among plants. Some have adhesive eggs. The mortality of the young and especially of the eggs is very high, and often only a few individuals grow to maturity out of hundreds, thousands, and in some cases millions of eggs laid.

Males produce sperm, usually as a milky white substance called milt, in two (sometimes one) testes within the body cavity. In bony fishes a sperm duct leads from each testis to a urogenital opening behind the vent or anus. In sharks and rays and in cyclostomes the duct leads to a cloaca. Sometimes the pelvic fins are modified to help transmit the milt to the eggs at the female's vent or on the substrate where the female has placed them. Sometimes accessory organs are used to fertilize females internally—for example, the claspers of many sharks and rays.

In the females the eggs are formed in two ovaries (sometimes only one) and pass through the ovaries to the urogenital opening and to the outside. In some fishes the eggs are fertilized internally but shed before development takes place. Members of about a dozen families each of bony fishes (teleosts) and sharks bear live young. Many skates and rays also bear live young. In some bony fishes the eggs simply develop within the female, the young emerging when the eggs hatch (ovoviviparous). Others develop within the ovary and are nourished by ovarian tissues after hatching (viviparous). There are also other methods utilized by fishes to nourish young within the female. In all live-bearers the young are born at a relatively large size and are few in number. In one family of primarily marine fishes, the surfperches from the Pacific coast of North America, Japan, and Korea, the males of at least one species appear to be born sexually mature, although they are not fully grown.

Hermaphroditism

Some fishes are hermaphroditic, an individual producing both sperm and eggs, usually at different stages of its life. Self-fertilization, however, is probably rare.

Successful reproduction and in many cases defense of the eggs and young is assured by rather stereotyped but often elaborate courtship and parental behaviour, either by the male or the female or both. Some fishes prepare nests by hollowing out depressions in the sand bottom (cichlids, for example), build nests with plant materials and sticky threads excreted by the kidneys (sticklebacks), or blow a cluster of mucus-covered bubbles at the water surface (gouramis). The eggs are laid in these structures. Some varieties of cichlids and catfishes incubate eggs in their mouths.

Some fishes, such as salmon, undergo long migrations from the ocean and up large rivers to spawn in gravel beds where they themselves hatched (anadromous fishes). Others undertake shorter migrations from lakes into streams or in other ways enter for spawning habitats that they do not ordinarily occupy.

FORM AND FUNCTION

Body plan. The basic structure and function of the fish body is similar to those of all other vertebrates. The usual four types of tissues are present: surface or epithelial, connective (bone, cartilage, and fibrous tissues, as well as their derivative, blood), nerve, and muscle tissues. In addition, the organs and organ systems parallel those of other vertebrates.

The typical fish body is streamlined and spindle-shaped, with an anterior head, gill apparatus, and heart, the latter lying in the midline just below the gill chamber (see Figure 1). The body cavity, containing the vital organs, is situated behind the head in the lower anterior part of the body. The anus usually marks the posterior termination of the body cavity and most often occurs just in front of the

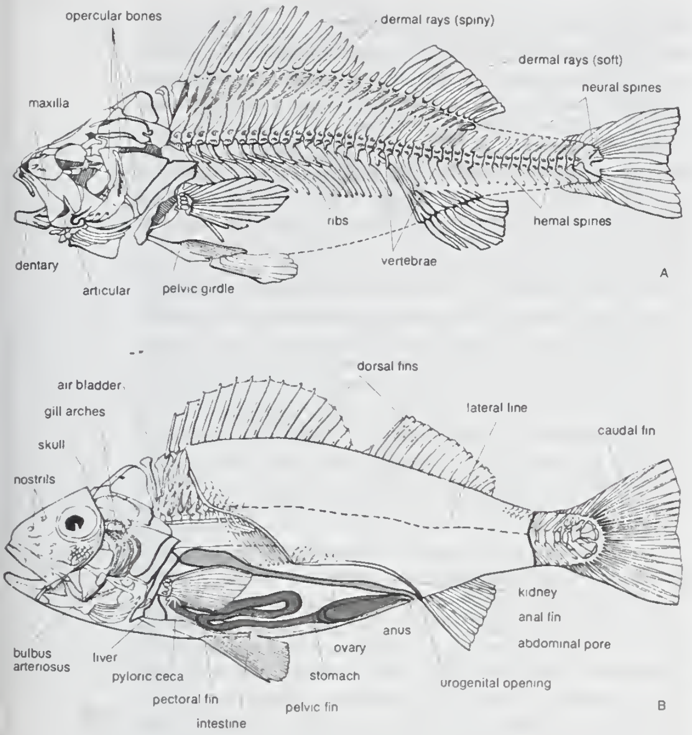


Figure 1: *Internal structure of fishes.*
(A) Skeleton of a perch. (B) Dissection of perch.

base of the anal fin. The spinal cord and vertebral column continue from the posterior part of the head to the base of the tail fin, passing dorsal to the body cavity and through the caudal (tail) region behind the body cavity. Most of the body is of muscular tissue, a high proportion of which is necessitated by swimming. In the course of evolution this basic body plan has been modified repeatedly into the many varieties of fish shapes that exist today.

Skeleton

The skeleton forms an integral part of the fish's locomotion system, as well as serving to protect vital parts. The internal skeleton consists of the skull bones (except for the roofing bones of the head, which are really part of the external skeleton), vertebral column, and the fin supports (fin rays). The fin supports are derived from the external skeleton but will be treated here because of their close functional relationship to the internal skeleton. The internal skeleton of cyclostomes, sharks, and rays is of cartilage; that of many fossil groups and some primitive living fishes is mostly of cartilage but may include some bone. In place of the vertebral column, the earliest vertebrates had a fully developed notochord, a flexible stiff rod of viscous cells surrounded by a strong fibrous sheath. During the evolution of modern fishes the rod was replaced in part by cartilage and then by ossified cartilage. Sharks and rays retain a cartilaginous vertebral column; bony fishes have spool-shaped vertebrae that in the more primitive living forms only partially replace the notochord. The skull, including the gill arches and jaws of bony fishes, is fully, or at least partially, ossified. That of sharks and rays remains cartilaginous, at times partially replaced by calcium deposits but never by true bone.

The supportive elements of the fins (basal or radial bones or both) have changed greatly during fish evolution. Some of these changes are described in the sections below (*Evolution and paleontology; Classification*). Most fishes possess a single dorsal fin on the midline of the back. Many have two and a few have three dorsal fins. The other fins are the single tail and anal fins and paired pelvic and pectoral fins. A small fin, the adipose fin, almost always without fin rays, occurs in many of the relatively primitive teleosts (such as trout) on the back near the base of the caudal fin.

The skin. The skin of a fish must serve many functions. It aids in maintaining the osmotic balance, provides physical protection for the body, is the site of coloration, contains sensory receptors, and, in some fishes, functions in respiration. Mucous glands, which aid in maintaining the water balance and offer protection from bacteria, are extremely numerous in fish skin, especially in cyclostomes and teleosts. Since mucous glands are present in the modern lampreys it is reasonable to assume that they were present in primitive fishes, such as the ancient Silurian and Devonian agnaths. Protection from abrasion and predation is another function of the fish skin, and dermal (skin) bone arose early in fish evolution in response to this need. It is thought that bone first evolved in skin and only later invaded the cartilaginous areas of the fish's body, to provide additional support and protection. There is some argument as to which came first, cartilage or bone, and fossil evidence does not settle the question. In any event, dermal bone has played an important part in fish evolution and has different characteristics in different groups of fishes. Several groups are characterized at least in part by the kind of bony scales they possess.

Types of scales

Scales have played an important part in the evolution of fishes. Primitive fishes usually had thick bony plates or thick scales in several layers of bone, enamel, and related substances. Modern teleost fishes have scales of bone, which, while still protective, allow much more freedom of motion in the body. A few modern teleosts (some catfishes, sticklebacks, and others) have secondarily acquired bony plates in the skin. Modern and early sharks possessed placoid scales, a relatively primitive type of scale with a toothlike structure, consisting of an outside layer of enamel-like substance (vitrodentine), an inner layer of dentine, and a pulp cavity containing nerves and blood vessels. Primitive bony fishes had thick scales of either the ganoid or the cosmoid type. Cosmoid scales have a hard, enamel-like outer layer, an inner layer of cosmine (a form of dentine), and then a layer of vascular bone (isopedine). In ganoid scales the hard outer layer is different chemically and is called ganoin. Under this is a cosmine-like layer and then a vascular bony layer. The thin, translucent bony scales of modern fishes, called cycloid and ctenoid scales (the latter distinguished by serrations at the edges), lack enameloid and dentine layers.

Skin has several other functions in fishes. It is well supplied with nerve endings and presumably receives tactile, thermal, and pain stimuli. Skin is well supplied with blood vessels. Some fishes breathe in part through the skin, by the exchange of oxygen and carbon dioxide between the surrounding water and numerous small blood vessels near the skin surface.

Coloration

Skin serves as protection through the control of coloration. Fishes exhibit an almost limitless range of colours. The colours often blend closely with the surroundings, effectively hiding the animal. Many fishes use bright colours for territorial advertisement or as recognition marks for other members of their own species, or sometimes for members of other species. Many fishes can change their colour to a greater or lesser degree, by expansion and con-

From T Storer and R Usinger *General Zoology* (1957) McGraw-Hill Book Company

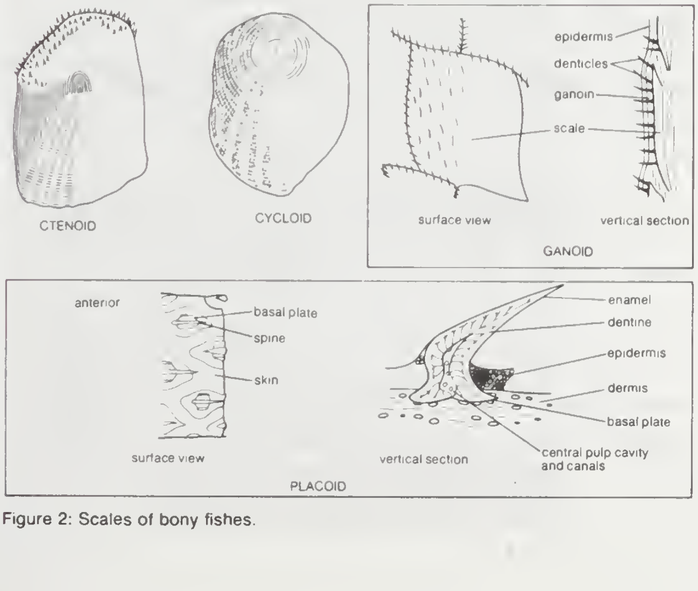


Figure 2: Scales of bony fishes.

traction of the pigment cells (chromatophores). Black pigment cells (melanophores), of almost universal occurrence in fishes, are often juxtaposed with other pigment cells. When placed near iridocytes or leucophores (bearing the silvery or white pigment guanine) melanophores produce structural colours of blue and green. These colours are often extremely intense, because they are formed by refraction of light through the needlelike crystals of guanine. The blue and green refracted colours are often relatively pure, lacking the red and yellow rays, which have been absorbed by the black pigment (melanin) of the melanophores. Yellow, orange, and red colours are produced by erythrophores, cells containing the appropriate carotenoid pigments. Other colours are produced by combinations of melanophores, erythrophores, and iridocytes.

The muscle system. The major portion of the body of most fishes consists of muscles. Most of the mass is trunk musculature, the fin muscles usually being relatively small. The caudal fin is usually the most powerful fin, with the largest amount of direct musculature. Its musculature is really a structural and functional continuation of the main musculature of the body. The body musculature is usually arranged in two rows of chevron-shaped segments on each side. Contractions of these segments, each attached to adjacent vertebrae and vertebral processes, bends the body on the vertebral joint, producing successive undulations of the body, passing from the head to the tail, and producing driving strokes of the tail. It is the latter that provides the strong forward movement for most fishes.

The digestive system. The digestive system, in a functional sense, starts at the mouth, with the teeth used to capture prey or collect plant foods. Mouth shape and tooth structure vary greatly in fishes, depending on the kind of food normally eaten. Most fishes are predacious, feeding on small invertebrates or other fishes and have simple conical teeth on the jaws, on at least some of the bones of the roof of the mouth, and on special gill arch structures just in front of the esophagus. The latter are throat teeth.

From B Dean, *Fishes Living and Fossil*, in A S Romer *The Vertebrate Body*, 4th ed (1970), W B Saunders Company

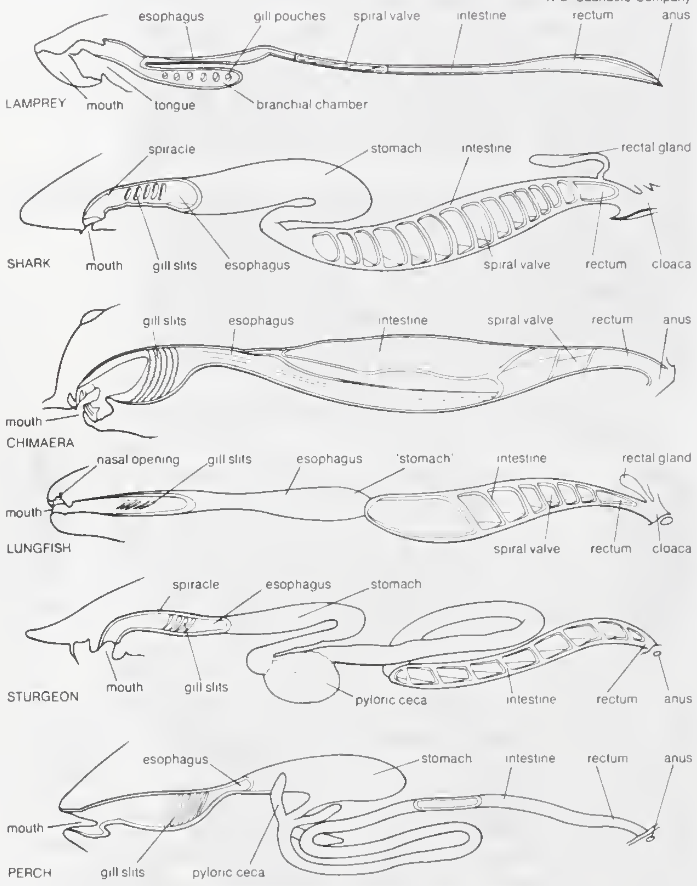


Figure 3: *Digestive tracts of various kinds of fishes.* Lampreys and chimaeras have no stomachs; the stomach of a lungfish is merely an enlargement of the esophagus.

Types of teeth

Most predacious fishes swallow their prey whole, and the teeth are used for grasping and holding prey, for orienting prey to be swallowed (head first) and for working the prey toward the esophagus. There are a variety of tooth types in fishes. Some, such as sharks and the piranhas, have cutting teeth for biting chunks out of their victims. A shark's tooth, although superficially like that of a piranha, appears in many respects to be a modified scale, while that of the piranha is like that of other bony fishes, consisting of dentine and enamel. Parrotfishes have beaklike mouths with short incisor-like teeth for breaking off coral and have heavy pavement-like throat teeth for crushing the coral. Some catfishes have small brushlike teeth, arranged in rows on the jaws, for scraping plant and animal growth from rocks. Many fishes (*e.g.*, the Cyprinidae or minnows) have no jaw teeth at all but have very strong throat teeth.

Some fishes gather planktonic food by straining it from their gill cavities with numerous elongate stiff rods (gill rakers), anchored by one end to the gill bars. The food collected on these rods is passed to the throat where it is swallowed. Most fishes have only short gill rakers that help keep food particles from escaping out the mouth cavity into the gill chamber.

Once reaching the throat, food enters a short, often greatly distensible esophagus, a simple tube with a muscular wall leading into a stomach. The stomach varies greatly in fishes, depending upon the diet. In most predacious fishes it is a simple straight or curved tube or pouch with a muscular wall and a glandular lining. Food is largely digested here and leaves the stomach in liquid form.

Between the stomach and the intestine, ducts enter the digestive tube from the liver and pancreas. The liver is a large, clearly defined organ. The pancreas may be imbedded in it, diffused through it, or broken into small parts spread along some of the intestine. The junction between the stomach and the intestine is marked by a muscular valve. Pyloric ceca (blind sacs) occur in some fishes at this junction and have a digestive or an absorptive function, or both.

The intestine itself is quite variable in length depending upon the diet. It is short in predacious forms, sometimes no longer than the body cavity, but long in herbivorous forms, being coiled and several times longer than the entire length of the fish in some species of South American catfishes. The intestine is primarily an organ for absorbing nutrients into the bloodstream. The larger its internal surface, the greater its absorptive efficiency, and a spiral valve is one method of increasing its absorption surface. Sharks, rays, chimaeras, lungfishes, surviving chondrosteans, holosteans, and even a few of the more primitive teleosts have a spiral valve or at least traces of it in the intestine. Most modern teleosts have increased the area of the intestinal walls by having numerous folds and villi (fingerlike projections) somewhat like those in man. Undigested substances are passed to the exterior through the anus in most teleost fishes. In lungfishes, sharks, and rays it is first passed through the cloaca, a common cavity receiving the intestinal opening and the ducts from the uro-genital system.

The respiratory system. Oxygen and carbon dioxide dissolve in water and most fishes exchange dissolved oxygen and carbon dioxide in water by means of the gills. The gills lie behind and to the side of the mouth cavity and consist of fleshy filaments supported by the gill arches and filled with blood vessels, which give gills a bright red colour. Water taken in continuously through the mouth passes backward between the gill bars and over the gill filaments, where the exchange of gases takes place. The gills are protected by a gill cover in teleosts and many other fishes, but by flaps of skin in sharks, rays, and some of the older fossil fish groups. The blood capillaries in the gill filaments are close to the gill surface to take up oxygen from the water and to give up excess carbon dioxide to the water.

Swim bladder

Most modern fishes have a hydrostatic (ballast) organ, called the swim bladder, that lies in the body cavity just below the kidney and above the stomach and intestine. It originated as a diverticulum of the digestive canal. In advanced teleosts, especially the acanthopterygians, the bladder has lost its connection with the digestive tract, a condition called physoclistic. The connection has been retained (physostomous) by many relatively primitive teleosts. In several unrelated lines of fishes the bladder has become specialized as a lung or, at least, as a highly vascularized accessory breathing organ. Some fishes with such accessory organs are obligate air breathers and will drown if denied access to the surface, even in well-oxygenated water. Fishes with a hydrostatic form of swim bladder can control their depth by regulating the amount of gas in the bladder. The gas, mostly oxygen, is secreted into the bladder by special glands, rendering the fish more buoyant; it is absorbed into the bloodstream by another special organ, reducing the overall buoyancy and allowing the fish to sink. Some deep-sea fishes may have oil in the bladder, rather than gas. Other deep-sea and some bottom-living forms have much reduced swim bladders or have lost the organ entirely.

The swim bladder of fishes follows the same developmental pattern as the lungs of land vertebrates. There is no doubt that the two structures have the same historical origin in primitive fishes. More or less intermediate forms still survive among the more primitive types of fishes such as the lungfishes *Lepidosiren* and *Protopterus*.

The circulatory system. The circulatory, or blood vascular, system consists of the heart, the arteries, the capillaries, and the veins: it is in the capillaries that the interchange of oxygen, carbon dioxide, nutrients, and other substances such as hormones and waste products takes place. The capillaries in turn lead to the veins, which return the venous blood with its waste products to the heart, kidneys, and gills. There are two kinds of capillary beds, those in the gills and those in the rest of the body. The heart, a folded continuous muscular tube with three or four sacklike enlargements, undergoes rhythmic contractions, and receives venous blood in a sinus venosus. It then passes the blood to an auricle and then into a thick, muscular pump, the ventricle. From the ventricle the blood goes to a bulbous structure at the base of a ventral aorta just below the gills. The blood then passes to the afferent (receiving) arteries of the gill arches and then to the gill capillaries. There waste gases are given off to the environment and oxygen is absorbed. From there the oxygenated blood enters efferent (exuant) arteries of the gill arches and then into the dorsal aorta. From there blood is distributed to the tissues and organs of the body. One-way valves prevent backflow. The circulation of fishes thus differs from that of the reptiles, birds, and mammals, in that oxygenated blood is not returned to the heart prior to distribution to the other parts of the body.

Excretory organs. The primary excretory organ in fishes, as in other vertebrates, is the kidney. In fishes some excretion also takes place in the digestive tract, skin, and especially the gills (where ammonia is given off). Compared with land vertebrates, fishes have a special problem in maintaining their internal environment at a constant concentration of water and dissolved substances, such as salts. Proper balance of the internal environment (homeostasis) of a fish is in a great part maintained by the excretory system, especially the kidney.

Water balance

The kidney, gills, and skin play an important role in maintaining a fish's internal environment and checking the effects of osmosis. Marine fishes live in an environment in which the water around them has a greater concentration of salts than they can have inside their body and still maintain life. Freshwater fishes, on the other hand, live in water with a much lower concentration of salts than they require inside their bodies. Osmosis tends to promote the loss of water from the body of a marine fish and absorption of water by that of a freshwater fish. Mucus in the skin tends to slow the process but is not a sufficient barrier to prevent the movement of fluids through the permeable skin. When solutions on two sides of a permeable membrane have different concentrations of dissolved substances, water will pass through the membrane into the more concentrated solution, while the dissolved chemicals move into the area of lower concentration (diffusion).

The kidney of freshwater fishes is often larger in relation to body weight than that of marine fishes. In both groups the kidney excretes wastes from the body, but that of freshwater fishes also excretes large amounts of water, counteracting the water absorbed through the skin. Freshwater fishes tend to lose salt to the environment and must replace it. They get some salt from their food, but the gills and skin inside the mouth actively absorb salt from water passed through the mouth. This absorption is performed by special cells capable (like those of the kidney) of moving salts against the diffusion gradient. Freshwater fishes drink very little water and take in little water in their food.

Marine fishes must conserve water, therefore their kidneys excrete little water. To maintain their water balance marine fishes drink large quantities of seawater, retaining most of the water and excreting the salt. By reabsorption of needed water in the kidney tubules, they discharge a more concentrated urine than do freshwater fishes. Most nitrogenous waste in marine fishes appears to be secreted by the gills as ammonia. Some marine fishes, at least, can excrete salt by clusters of special cells in the gills and intestine.

There are several teleosts—for example, the salmon—that travel between fresh water and seawater and must adjust to the reversal of osmotic gradients. They adjust their physiological processes by spending time (often surprisingly little time) in the intermediate brackish environment.

Marine lampreys, hagfishes, sharks, and rays have osmotic concentrations in their blood about equal to that of seawater so do not have to drink water nor perform much physiological work to maintain their osmotic balance. In sharks and rays the osmotic concentration is kept high by retention of urea in the blood. Freshwater sharks have a lowered concentration of urea in the blood.

Endocrine glands. Endocrine glands secrete their products into the bloodstream and body tissues and, along with the central nervous system, control and regulate many kinds of body functions. Cyclostomes have a well-developed endocrine system, and presumably it was well developed in the early Agnatha, ancestral to modern fishes. Although the endocrine system in fishes is similar to that of higher vertebrates, there are numerous differences in detail. The endocrine glands of fishes are the pituitary, thyroid, suprarenals, adrenals, pancreatic islets, sex glands (ovaries and testes), the inner wall of the intestine, and the ultimobranchial bodies. There are some others whose function is not well understood. These organs regulate sexual activity and reproduction, growth, osmotic pressure, general metabolic activities such as the storage of fat and the utilization of foodstuffs, blood pressure, and certain aspects of skin colour. Many of these activities also are controlled in part by the central nervous system, which works with the endocrine system in maintaining the life of a fish. Some parts of the endocrine system are developmentally, and undoubtedly evolutionarily, derived from the nervous system.

The nervous system and sensory organs. As in all vertebrates, the nervous system of fishes is the primary mechanism coordinating body activities, as well as integrating these activities in the appropriate manner with stimuli from the environment. The central nervous system, the brain, and spinal cord, are the primary integrating mechanisms. The peripheral nervous system, consisting of nerves that connect the brain and spinal cord to various body organs, carries sensory information from special receptor organs such as the eyes, internal ears, nares (sense of smell), taste glands, and others to the integrating centres of the brain and spinal cord. The peripheral nervous system also carries information via different nerve cells from the integrating centres of the brain and spinal cord. This coded information is carried to the various organs and body systems, such as the skeletal muscular system, for appropriate action in response to the original external or internal stimulus. Another branch of the nervous system, the autonomic system, helps to coordinate the activities of many glands and organs and is itself closely connected to the integrating centres of the brain.

The brain

The brain of the fish is divided into several anatomical and functional parts, all closely interconnected but each serving as the primary centre of integrating particular kinds of responses and activities. Several of these centres or parts are primarily associated with one type of sensory perception such as sight, hearing, or smell (olfaction).

Olfaction. The sense of smell is important in almost all fishes. Certain eels with tiny eyes depend mostly on smell for location of food. The olfactory, or nasal, organ of fishes is located on the dorsal surface of the snout. The lining of the nasal organ has special sensory cells that perceive chemicals dissolved in the water such as substances from food material and send sensory information to the brain by way of the first cranial nerve. Odour also serves as an alarm system. Many fishes, especially various species of freshwater minnows, react with alarm to the body fluids produced by an injured member of their own species.

Taste. Many fishes have a well-developed sense of taste, and tiny pitlike taste buds or organs are located not only within their mouth cavities but also over their heads and parts of their body. The barbels ("whiskers") of catfishes, which often have poor vision, serve as supplementary taste organs, those around the mouth being actively used to search out food on the bottom. Some species of naturally blind cave fishes are especially well supplied with taste buds, these often covering most of their body's surface.

Sight. Sight is extremely important in most fishes. The eye of a fish is basically like that of all other vertebrates, but the eyes of fishes are extremely varied in structure and adaptation. In general, fishes living in dark and dim water habitats have large eyes, unless they have specialized in some compensatory way so that another sense (such as smell) is dominant, in which case the eyes will often be reduced. Fishes living in brightly lighted shallow waters often will have relatively small but efficient eyes. Cyclostomes have somewhat less elaborate eyes than other fishes, with skin stretched over the eyeball perhaps making their vision somewhat less effective. Most fishes have a spherical lens and accommodate their vision to far or near subjects by moving the lens within the eyeball. A few sharks accommodate by changing the shape of the lens, as in land vertebrates. Those fishes that are heavily dependent upon the eyes have especially strong muscles for accommodation. Most fishes see well, despite the restrictions imposed by frequent turbidity of the water and by light refraction. Experimental evidence indicates that many shallow-water fishes, if not all, have colour vision and see some colours especially well, but some bottom-dwelling shore fishes live in areas where the water is sufficiently deep to filter out most if not all colours, and these fishes apparently never see colours. When tested in shallow water, they apparently are unable to respond to colour differences.

Colour vision

Hearing. Sound perception and balance are intimately associated senses in a fish. The organs of hearing are entirely internal, located within the skull, on each side of the brain and somewhat behind the eyes. Sound waves, especially those of low frequencies, travel readily through water and impinge directly upon the bones and fluids of the head and body, to be transmitted to the hearing organs. Fishes readily respond to sound; for example, a trout conditioned to escape by the appoach of fishermen will take flight upon perceiving footsteps on a stream bank even if it cannot see the fisherman. Compared with humans, however, the range of sound frequencies heard by fishes is greatly restricted. It is thought that many fishes communicate with each other in a crude way by producing sounds in their swim bladders, in their throats by rasping their teeth, and in other ways.

Other senses (touch, pain, and special senses). A fish or other vertebrate seldom has to rely on a single type of sensory information to determine the nature of the environment around it. A catfish uses taste and touch when examining a food object with its oral barbels. Like most other animals, fishes have many touch receptors over their body surface. Pain and temperature receptors also are present in fishes and presumably produce the same kind of information to a fish as to humans. Fishes react in a negative fashion to stimuli that would be painful to human beings, suggesting that they feel a sensation of pain.

Lateral line system

An important sensory system in fishes that is absent in other vertebrates (except some amphibians) is the lateral line system. This consists of a series of heavily innervated small canals located in the skin and bone around the eyes, along the lower jaw, over the head and down the midside of the body where it is associated with the scales. Intermittently along these canals are located tiny sensory organs (pit organs) that apparently detect changes in pressure. The system allows a fish to sense changes in water currents and pressure, thereby helping the fish to orient itself to the various changes that occur in the physical environment.

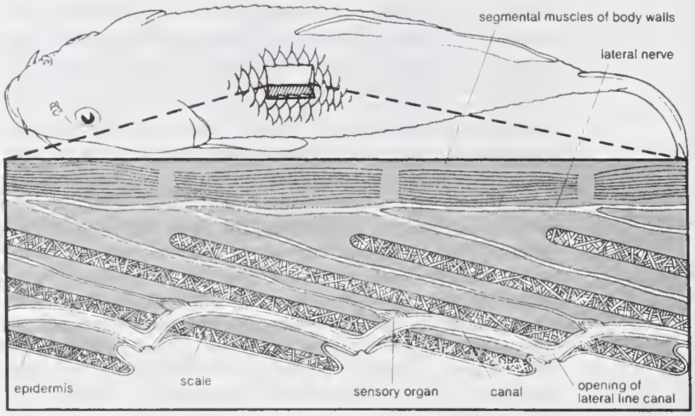

segmental muscles of body walls
lateral nerve
epidermis
scale
sensory organ
canal
opening of lateral line canal

Figure 4. Magnified longitudinal section through a carp's body wall showing lateral line sensory system.
From T. Storer and R. Usinger, *General Zoology* (1957); McGraw-Hill Book Co

EVOLUTION AND PALEONTOLOGY

Although a great many fossil fishes have been found and described, they represent a tiny portion of the long and complex evolution of fishes and knowledge of fish evolution remains relatively fragmentary. In the classification presented in this article fishlike vertebrates are divided into seven classes, the members of each having a different basic structural organization and different physical and physiological adaptations for the problems presented by the environment. The broad basic pattern has been one of successive replacement of older groups by newer, better adapted groups. One or a few members of a group evolved a basically more efficient means of feeding, breathing, swimming, or several better ways of living. These better adapted groups then forced the extinction of members of the older group with which they competed for available food, breeding places, or other necessities of life. As the new fishes became well established, some of them evolved further and adapted to other habitats, where they continued to replace members of the old group already there. The process was repeated until all or almost all members of the old group in a variety of habitats had been replaced by members of the newer evolutionary line.

Early jawless fishes

The earliest vertebrate fossils of certain relationships are fragments of dermal armour of jawless fishes (class Agnatha, order Heterostraci) from the Middle Ordovician Period in North America, about 450,000,000 years in age. Early Ordovician toothlike fragments from the U.S.S.R. are less certainly remains of the class Agnatha. It is uncertain whether the North American jawless fishes inhabited shallow coastal marine waters, where their remains became fossilized, or were freshwater vertebrates washed into coastal deposits by stream action.

Jawless fishes probably arose from ancient small, soft-bodied filter-feeding organisms much like and probably also ancestral to the modern sand-dwelling filter feeders, the Cephalochordata (*Amphioxus* and its relatives). The body in the ancestral animals was probably stiffened by a notochord. Although a vertebrate origin in fresh water is much debated by paleontologists, it is possible that mobility of the body and protection provided by dermal armour arose in response to streamflow in the freshwater environment and to the need to escape from and resist the clawed invertebrate eurypterids that lived in the same waters. Because of the marine distribution of the surviving primitive chordates, many paleontologists doubt that the vertebrates arose in fresh water.

Heterostracan remains are next found in what appear to be delta deposits in two North American localities of Silurian age. By the close of the Silurian, about 400,000,000 years ago, European heterostracan remains are found in what appear to be delta or coastal deposits. In the Late Silurian of the Baltic area, lagoon or freshwater deposits yield jawless fishes of the order Osteostraci. Somewhat later in the Silurian from the same region, layers contain fragments of jawed acanthodians, the earliest group of jawed vertebrates, and of jawless fishes. These layers lie between marine beds but appear to be washed out from fresh waters of a coastal region.

It is evident, therefore, that by the end of the Silurian both jawed and jawless vertebrates were well established and already must have had a long history of development. Yet paleontologists have remains only of specialized forms that cannot have been the ancestors of the placoderms and bony fishes that appear in the next period, the Devonian. No fossils are known of the more primitive ancestors of the agnaths and acanthodians. The extensive marine beds of the Silurian and those of the Ordovician are essentially void of vertebrate history. It is believed that the ancestors of fishlike vertebrates evolved in upland fresh waters, where whatever few and relatively small fossil beds were made probably have been long since eroded away. Remains of the earliest vertebrates may never be found.

By the close of the Silurian, all five known orders of jawless vertebrates had evolved, except perhaps the modern cyclostomes, which are without the hard parts that ordinarily are preserved as fossils. Cyclostomes were unknown as fossils until 1968, when a lamprey of modern body structure was reported from the Middle Pennsylvanian of Illinois, in deposits almost 300,000,000 years old. Fossil evidence of the four orders of armoured jawless vertebrates is absent from deposits later than the Devonian. Presumably they became extinct at that time, being replaced by the more efficient and probably more aggressive placoderms, acanthodians, selachians (sharks and relatives), and by early bony fishes. Cyclostomes survived probably because they early evolved from anaspid agnaths and developed a rasping tonguelike structure and a sucking mouth, enabling them to prey on other fishes. With this way of life they apparently had no competition from other fish groups.

Early jawless vertebrates probably fed on tiny organisms by filter feeding, as do the larvae of their descendants, the modern lampreys. The gill cavity of the early agnaths was large. It is thought that small organisms taken from the bottom by a nibbling action of the mouth, or more certainly by a sucking action through the mouth, were passed into the gill cavity along with water for breathing. Small organisms then were strained out by the gill apparatus and directed to the food canal. The gill apparatus thus evolved as a feeding, as well as a breathing, structure. The head and gills in the agnaths were protected by a heavy dermal armour; the tail region was free, allowing motion for swimming.

Appearance of bone

Most important for the evolution of fishes and vertebrates in general was the early appearance of bone, cartilage, and enamel-like substance. These materials became modified in later fishes, enabling them to adapt to many aquatic environments and finally even to land. Other basic organs and tissues of the vertebrates such as the central nervous system, heart, liver, digestive tract, kidney, and circulatory system undoubtedly were present in the ancestors of the Agnatha. In many ways, bone, both external and internal, was the key to vertebrate evolution.

The next class of fishes to appear was the Acanthodii, containing the earliest known jawed vertebrates, which arose in the Upper Silurian, over 400,000,000 years ago. The acanthodians declined after the Devonian but lasted into the Lower Permian, a little less than 280,000,000 years ago. The first complete specimens appear in Lower Devonian freshwater deposits, but later in the Devonian and Permian some members appear to have been marine. Most were small fishes, not over 75 centimetres (approximately 30 inches) in length.

We know nothing of the ancestors of the acanthodians. They must have arisen from some jawless vertebrate, probably in fresh water. They appear to have been active swimmers with almost no head armour but with large eyes, indicating that they depended heavily on vision. Perhaps they preyed on invertebrates. The rows of spines and spinelike fins between the pectoral and pelvic fins give some credence to the idea that paired fins arose from "fin folds" along the body sides.

The relationships of the acanthodians to other jawed vertebrates are obscure. They possess features found in both sharks and bony fishes. They are like early bony fishes in possessing ganoid-like scales and a partially ossified internal skeleton. Certain aspects of the jaw appear to be more like those of bony fishes than sharks, but the bony fin spines and certain aspects of the gill apparatus would seem to favour relationships with early sharks. Acanthodians do not seem particularly close to the Placodermi although, like the placoderms, they apparently possessed less efficient tooth replacement and tooth structure than the sharks and the bony fishes, possibly one reason for their subsequent extinction.

The first record of the jawed Placodermi is from the Early Devonian, about 390,000,000 years ago. The placoderms flourished for about 60,000,000 years and were almost gone at the end of the Devonian. Nothing is known of their ancestors, who must have existed in the Silurian. The evolution of several other, better adapted, fish groups soon followed the appearance of the placoderms and this apparently led to their early extinction. Their greatest period of success was approximately during the middle of the Devonian, when some of them became marine. As their name indicates (placoderm means "plate skin"), most of these fishes had heavy coats of bony armour, especially about the head and anterior part of the body. The tail remained free and heterocercal (*i.e.*, the upper lobe long, the lower one small or lacking). Most placoderms remained small, 30 centimetres or less in length, but one group, the arthrodires, had a few marine members that reached 10 metres in length. Important evolutionary advances of the placoderms were in the jaws (which usually were amphistylic—*i.e.*, involving the hyoid and quadrate bones) and development of fins, especially the paired fins with well-formed basal or radial elements. The jaws tended to be of single elements with strongly attached toothlike structures. These were too specialized to be considered ancestral to the more adaptable jaws of subsequent bony fish groups. It has been proposed that sharks arose from some group of placoderms near the Stensioelliformes and that the chimaera line (class Holocephali) arose from certain arthrodires; this suggestion, however, is uncertain.

A peculiar, five-centimetre fish, *Palaeospondylus*, from Middle Devonian rocks in Scotland, is probably not a placoderm, although classed with them here. Various suggestions that its relationships are with the agnaths, placoderms, acanthodians, sharks, and even lungfishes and amphibians are unconvincing and its relationships remain completely unknown.

Origin of sharks

Sharks (class Selachii) first appear in the Middle Devonian about 375,000,000 years ago, became quite prominent by the end of the Devonian, and are still successful today. Two Early Devonian orders of primitive sharklike fishes, the Cladoselachiformes and the Cladodontiformes, became extinct by the end of the Permian, about 230,000,000 years ago, while the freshwater order Xenacanthiformes lasted until the Middle Triassic, about 200,000,000 years ago. The final Devonian order, Heterodontiformes, still has surviving members.

Modern sharks and rays arose during the Jurassic Period, about 135,000,000 to 190,000,000 years ago, probably from an older group, the hybodont sharks. Presumably marine cladoselachians gave rise to the hybodont Heterodontiformes during the close of the Devonian. These had the placoderm amphystylic jaws but had paired fins of a more efficient type. In turn the hybodonts are thought to have given rise to the living but archaic mollusk-eating Port Jackson sharks (heterodonts). The relationships of the surviving (but archaic) hexanchiform sharks are unknown. The two main orders of modern Selachii, the Lamniformes (sharks) and Rajiformes (skates and rays), appeared between 140,000,000 and 180,000,000 years ago during the Jurassic Period. They are characterized by a hyostylic jaw (in which articulation involves only the hyoid bone), an

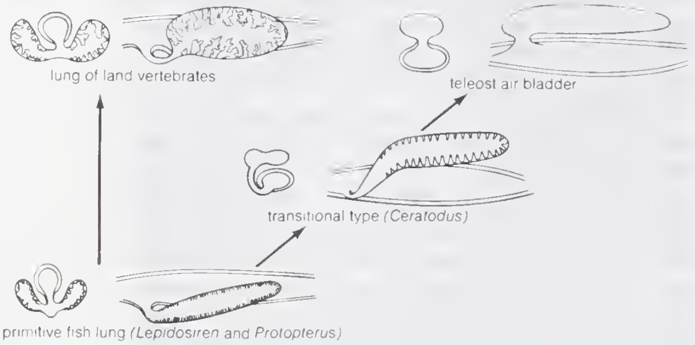

lung of land vertebrates
teleost air bladder
transitional type (*Ceratodus*)
primitive fish lung (*Lepidosiren* and *Protopterus*)

Figure 5: *Evolution of lungs and swim bladder.*
Each example consists of (left) a cross section through the gut and saclike growth and (right) a longitudinal section.

From A.S. Romer *The Vertebrate Story*, © 1933, 1939, 1941 1959 by the University of Chicago, all rights reserved

improvement allowing greater mobility of the jaws and an important feature in the methods of predation used by modern selachians.

Skates and rays evolved from some bottom-living shark-like ancestor during the Jurassic. The primary evolution and diversification of modern sharks, skates, and rays took place in the Cretaceous Period and Cenozoic Era. Thus, along with the teleost fishes (discussed below) most surviving sharks, skates, and rays are essentially of relatively recent origin, their main evolutionary radiation having taken place within the last 140,000,000 years.

The class Holocephali, the chimaeras or ratfishes as their modern survivors are called, first appeared in the Upper Devonian but were most common and diversified during the Mesozoic Era. Only one of the seven known orders survived beyond the close of the Cretaccous Period about 65,000,000 years ago. Although not many modern species of chimaeras are known, they are sometimes relatively abundant in their deep-sea habitat. The relationships of these fishes are in question. It has been proposed that they are related to the Devonian ptyctodont arthrodires, which had a chimaera-like shape and pelvic claspers. It has also been suggested that they are closely related to the Selachii because both selachians and holocephalians have many characters in common, such as placoid scales, pelvic claspers, and absence of true bone. It has been suggested recently that both holocephalians and selachians are related to the acanthodians on the basis of the gill arch structures. Further evidence is needed to solve the problem of their classification and relationships.

The class Sarcopterygii are extremely ancient in origin, their first remains appearing in Lower Devonian strata of Germany, about 390,000,000 years old. The most important group, the rhipidistians, which gave rise to the amphibians by the end of the Devonian, became extinct about 120,000,000 years later, near the beginning of the Permian. Two lesser groups, the coelacanths and the dipnoans (lungfishes), have barely survived. Recognition of the class Sarcopterygii is controversial in that some ichthyologists believe that the two major groups, the Crossopterygii (including the rhipidistians and coelacanths) and the Dipnoi, have arisen from independent origins, the present structures of the two groups being widely divergent. The primitive members, however, show several similarities, supporting the view that they had a common ancestor. The nature of the ancestor or ancestors remains a mystery. The Sarcopterygii probably evolved from unknown Silurian jawed freshwater fishes that may also have been ancestral to the actinopterygians.

The rhipidistian crossopterygians apparently flourished in the fresh waters of the Middle Devonian where, in adapting to a habitat subject to seasonal droughts, some evolved pectoral and pelvic appendages strong enough and flexible enough to enable them to leave drying pools to seek out those ponds that retained water. Paradoxically, terrestrial amphibians first rose through the need to survive in water.

The early coelacanths of the Upper Devonian were small freshwater and inshore fishes, and it was not until the Late Permian and Triassic that they became marine and grew larger and more diverse. They are not known as fossils later than the Cretaceous, and it was therefore a great surprise when in 1938 a live, 160-centimetre (63-inch) specimen was taken at 120 metres (approximately 390 feet) off the coast of eastern South Africa.

The dipnoans first appeared in the Lower Devonian and were fully differentiated at that time. They flourished until the close of the Triassic, when their numbers became greatly reduced. The modern Australian lungfish differs little from one of the Triassic forms. The living South American and especially African lungfishes are elongated, specialized fishes adapted to live and survive in more or less annual ponds.

The Actinopterygii, or "ray-finned" fishes, is the largest class of fishes. In existence for about 390,000,000 years, since the Lower Devonian, it consists of some 52 orders containing more than 480 families, at least 80 of which are known only from fossils. The class contains the great majority of known living and fossil fishes, with about 20,000 living species. The history of actinopterygians can be divided into three basic stages or evolutionary radiations, each representing a different level of structural organization and efficiency.

The Chondrostei have a 300,000,000-year history. They Chondrosts

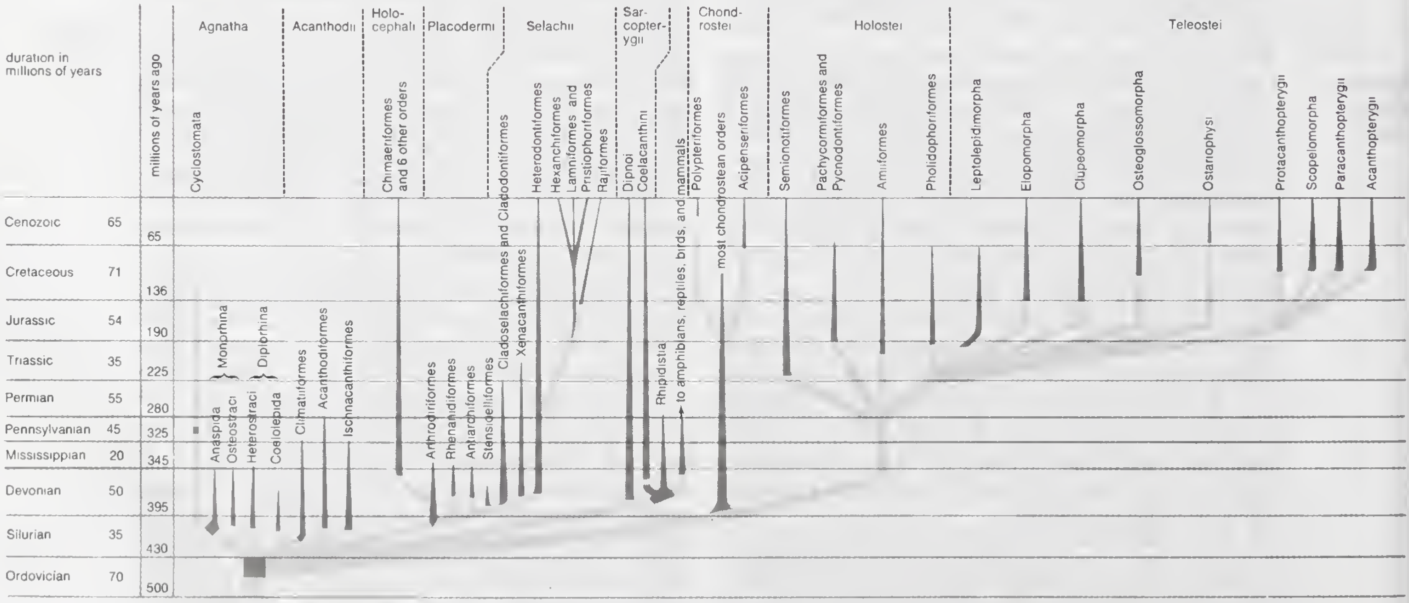


Figure 6: Phyletic family tree for fishes.

arose first in the Lower Devonian, increased in numbers and complexity until about the Permian, and thereafter declined, becoming almost extinct by the middle of the Cretaceous, 100,000,000 years ago. The chondrostean order Palaeonisciformes is the basal actinopterygian stock from which all other chondrosteans and the holosteans evolved. They were the most common fishes of their time, relatively small and typically like later fishes in appearance. In comparison with today's fishes they had peculiar looking jaws and tails. Their tails were heterocercal. On their bodies were thick ganoid scales that abutted each other, rather than overlapping as in most modern fishes. Palaeonisciformes often had large eyes placed far forward, long mouths with the upper jaw firmly bound to the fully armoured cheek, and a relatively weak lower jaw muscle. They gave rise to a great variety of types, with elongate bodies and jaws, bottom-living types that fed on microorganisms, deep-bodied marine reef fishes and coral-eating reef fishes. Almost all of these were replaced by modern teleosts. Surviving Chondrostei are the bottom-feeding marine and freshwater sturgeons, the strange plankton-feeding paddlefishes of the Mississippi of North America and the Yangtze River of China, and the freshwater bichirs and reed fishes (family Polypteridae) of Africa. The relationship of the polypterids is in some doubt, and the group is sometimes placed in the Sarcopterygii.

Several of the chondrostean orders developed characteristics that approached the holostean level of anatomic organization and are sometimes called subholosteans. One of these orders, the Parasemionotiformes, evolved from the Palaeonisciformes in the Lower Triassic and may have given rise to at least some of the holosteans. This evolutionary line leads to the Pholidophoriformes, which gave rise to modern bony fishes, or teleosts.

The holosteans are thought to be of mixed origin and to represent a stage in the evolution of a group of chondrostean orders. If so, the infraclass Holostei does not represent a single lineage. Important holostean characteristics are the approach of the tail toward the homocercal condition and the equal number of fin rays and basal elements of the fin rays. Both of these conditions make the holostean a more efficient swimmer than the chondrostean, as does thinning of the holostean body scales. Another important advance of holosteans was the freeing of the upper jaw from the preopercular bone of the cheek, allowing greater movement of the gill chamber and jaws, with more powerful development of the lower jaw muscle.

Five orders of holosteans are known, with their greatest evolutionary radiation occurring during the Triassic, Jurassic, and Cretaceous periods, when the chondrosteans were declining and the teleosts just beginning to expand. Two holostean groups survive today: the bowfin, *Amia calva*, and the several species of gars, *Lepisosteus*, all found in North America.

Advent of advanced bony fishes

The modern bony fishes, infraclass Teleostei, include the great majority of living fishes. They first appear in the fossil record about 190,000,000 years ago (as the family Leptolepididae) with their homocercal caudal fin and caudal skeleton already fully developed. They arose from an order of holosteans now extinct, the Pholidophoriformes. This group was intermediate in character between the chondrosteans and the teleosts. Teleosts have reached their fullest extent within the last 50,000,000 years and represent a distinct functional advance over their holostean ancestors. They have greater swimming ability, owing to the improvement in the tail structure, and have a still more efficient feeding and gill ventilating apparatus.

The bony fishes represent the culmination of a long evolution toward a body plan with maximum swimming efficiency. Particularly important in this evolution have been changes in fins and in the tail. Some authorities believe that the paired fins arose from a single continuous tail and anal fin that was divided at the vent and extended forward along each side to the head. Later the sections between the pectoral, pelvic, anal, and caudal fins were lost. The fin rays of sharks and rays are of a horny material, but those of many primitive fossil fishes are of bone. The bony fin rays of sarcopterygians and actinopterygians probably arose from scales lying in the fin folds. Modern teleost fishes have flexible fin rays (called soft rays) of jointed segments of bone, or spiny rays, each of solid continuous bone. The first dorsal fin of acanthopterygian fishes is of the spiny type. The original tail fin of primitive fishes was not an effective swimming organ, because of its asymmetry. The steady improvement in tail shape over 400,000,000 years is one of the prominent features of fish evolution. In primitive fishes the tail (vertebral) axis turned upward (heterocercal) or downward (hypocercal) and a lobe of flesh projected from it. This form of tail cannot provide a powerful driving mechanism, because the driving force is unevenly distributed relative to the body axis. With an asymmetrical tail, the fish swims by an undulating motion of the body and tail. In some fishes with a diphycercal tail (with the axis of the vertebrae extending down the middle of the fin lobe), developed in both modern and ancient fishes, the tail remains relatively ineffective because it has remained too rigid for proper propulsive action. The development of a true homocercal tail fin, in which powerful muscles move strong fin rays with a very flexible basal joint and in which the upper and lower lobes are about equal, is a development exclusive to teleost fishes.

Wide variation in form and habitat

As suggested by the existence of more than 400 families, teleosts are extremely varied in anatomical form and in the habitat occupied. They can be divided into about nine superorders, each with distinct evolutionary significance. The Leptolepidimorpha, an extinct, relatively primitive group, has uncertain relationships with other teleosts and is as yet poorly understood. The second group, the Elopomorpha, retains some relatively primitive living members, such as the tarpons, but is mostly represented by the large variety of specialized true eels. The Clupeomorpha includes the herrings and anchovies, relatively primitive fishes, mostly specialized for existence near the surface of the open ocean. A few species are anadromous, breeding in fresh water but spending most of their lives in the sea. The fourth superorder, the Osteoglossomorpha, consists of a group of relatively primitive teleosts, most of which are now extinct. The few surviving members are tropical and worldwide in distribution but adapted for restricted habitats. The Protacanthopterygii is a varied collection of relatively primitive orders, marine, deep-sea, and freshwater in distribution; trouts, smelts, and argentines are examples. The sixth superorder, the Ostariophysi, is an important group of primarily freshwater fishes, including the characins, carps, minnows, loaches, suckers, and catfishes.

The remaining three superorders have a complex fossil history and are not yet fully understood, but all seem to possess similar evolutionary trends. Each group shows a tendency to develop spiny fin rays in the dorsal and anal fins (reduced in some) and a shelf of bone under the eye. There is a tendency for the pelvic fins to move forward on the body, with a reorganization of swimming methods and a slight gain in manoeuvrability. All three groups probably are related and presumably arose from some early protacanthopterygian ancestor. The Scopelomorpha includes a wide variety of deep-sea open-ocean plankton feeders and predators, some of which bear light organs. The Paracanthopterygii is a rather miscellaneous collection of fishes, the most important to man being the cods. The final superorder, the Acanthopterygii, is the result of the great radiation of modern spiny-rayed fishes and contains the dominant fishes in marine shore habitats, tropical, temperate, and Arctic. They also have penetrated the freshwater environment, especially lakes, slow-moving streams, and ponds. The superorder has some important open-ocean members, such as tunas. The key to the successful acanthopterygian radiation probably has been their mobile, protractile mouth.

CLASSIFICATION

Distinguishing taxonomic features. In forming hypotheses about the evolution of fishes and in establishing classifications based on these hypotheses, ichthyologists place special emphasis on the comparative study of the skeleton. There are two primary advantages of this approach. First, direct comparison between extant and fossil groups is possible, the latter usually represented only by bony remains. The second advantage is that the bones of living fishes are relatively easy to observe and to study, compared with other body structures. Proper preservation and special preparation of the nervous system, for example, is difficult and expensive when the fishes compared are from the far ends of the earth. In the study of the relationships of species within a group major use has been made of similarities and differences in the dimensions of external features such as head and body length, and of counts of external characters, such as teeth, fin rays, and scales. Colour pattern is also important. In recent years valuable data on classification of fishes has been obtained from studies of comparative behaviour, physiology, genetics and functional anatomy.

Annotated classification. The following classification has been derived primarily from the works of C. Patterson, Miles, P.H. Greenwood and co-workers, D.E. Rosen and C. Patterson, and K.S. Thomson. Fish classification has undergone major revisions in recent years and further modifications can be expected in the future. Ichthyologists frequently disagree on major as well as minor concepts of phyletic relationships. There remains much to learn about both living and fossil fishes. The geographical distribution given for a poorly known fossil group usually represents only the location of fossil finds not necessarily the true distribution of the group. In the classification presented here groups indicated by a dagger (†) are known only from fossils.

CLASS AGNATHA

Vertebrates with a suctorial or filter-feeding mouth; no true jaws; 2 (possibly 1 sometimes) semicircular canals; pelvic fins lacking, pectoral finlike structures, when present, lacking fin rays; persistent notochord, without bone or cartilage; bony skeleton, when present, formed in skin; true gill arches absent, gill basket present. Habitat of fossil groups uncertain; earliest probably in fresh water.

Subclass Monorhina

With 1 nostril.

†*Order Osteostraci.* Late Silurian to close of Devonian. Heavily armoured with bony plates and scales; bony head shield present; bone cells tend to be absent; pectoral appendages present in some; eyes dorsal, close together; no common gill opening; bottom-dwelling with heterocercal tails. Length about 8–75 cm (roughly 3 to 30 in.).

†*Order Anaspida.* Late Silurian to Late Devonian. Heavily armoured with bony plates and scales; head protected by small bonelike plates; bone cells present; pectoral appendages various, spine or fleshy fold; eyes not close together but facing laterally; no common gill opening; probably swam above bottom with tail lobe extending downward (hypocercal). Length about 10–25 cm (roughly 4 to 10 in.).

Order Cyclostomata (Lampreys and hagfishes). Pennsylvanian and Recent. Freshwater and marine, breeding in fresh water (lampreys); or marine only (hagfishes). Without dermal ossification of any sort; pectoral appendages absent; eyes more or less lateral or dorsal (poorly developed in hagfishes); gill openings multiple, not common; tail more or less diphycercal. Primarily bottom-dwelling fishes, but suctorial, feeding on blood and juices of live fishes (lampreys) or rasping and feeding on flesh of dead or dying fishes (hagfishes); horny teeth present. Length about 15–100 cm (roughly 6 to 40 in.).

†Subclass Diplorhina

With 2 nostrils.

†*Order Heterostraci.* Ordovician to Upper Devonian. Usually heavily armoured, with a head shield in many species and scales or plates; bone cells absent; thin layer of enamel present over bone surface; no paired fins; eyes lateral and far apart; common gill opening present; tail hypocercal. Some perhaps midwater swimmers, others flattened bottom forms. Length about 5 to at least 30 cm (roughly 2 to 12 in.).

†*Order Coelolepida.* Upper Silurian and Lower Devonian. Small, little known agnaths of uncertain affinities. Appear to have had armoured head; body with small bony plates or scales; paired fins uncertain; eyes lateral; gill openings uncertain; number of nostrils uncertain; tail apparently reversed heterocercal. Length up to about 10 cm (roughly 4 in.).

†CLASS ACANTHODII

Jaws apparently formed of the 3rd gill arch (as in all jawed vertebrates) but attached to cranium and hyoid (4th) arch (amphistylic); jaws with teeth; pectoral and pelvic fins present, often with additional paired fins or spines between these; all fins except tail with a strong anterior spine; body covered with bony scales of the ganoid type, with bone cells tending to be lost in some members; no head shields but small dermal plates over head; some known with partially ossified vertebral column; cranium partially ossified; gill opening between jaws and hyoid arch apparently reduced to a spiracle; 3 semicircular canals (as in all higher fishes); an opercle (gill cover) present, attached to hyoid arch, followed by a series of smaller opercles, 1 over each remaining gill opening; caudal fin heterocercal. Active, free-swimming fishes with large eyes placed laterally. Small species probably freshwater, some larger species may have gone to sea.

†*Order Climatiiformes.* Upper Silurian to Lower Carboniferous. With 2 dorsal fins; 2 or more free spines present between pectoral and pelvic fins; operculum not covering entire gill chamber; supplementary operculae present or in some cases operculum covering entire gill chamber; no extramandibular bone. Mostly small, about 10 cm (roughly 4 in.) long.

†*Order Ischnacanthiformes.* Upper Silurian to Middle Carboniferous. With 2 dorsal fins; no free spines between pectoral and pelvic fins; operculum complete; no extramandibular bone.

†*Order Acanthodiformes.* Upper Silurian to Lower Permian. With 1 dorsal fin; no intermediate spines (or with a single pair) between pectoral and pelvic fins; operculum covering all or nearly all of opercular chamber; extramandibular bone present. Length to about 30 cm (roughly 12 in.).

†CLASS PLACODERMI (placoderms)

Jaws amphystylic (supported by both the cranium and hyoid arch); pelvic fins present or absent; pectoral fins or finlike structures often present; often with ossified or partially ossified vertebral column and internal cranium; gill arches present; skeletal ossification reduced in some groups; caudal fin most often heterocercal or some modification of this form. Habitat in many cases uncertain, but apparently most later groups were marine.

†*Order Arthrodiriformes* (arthrodires). Throughout the Devonian, especially common in last half of the period. Head and gill and trunk (thoracic) shields present (the two hinged upon each other in later Devonian forms); entire body more or less fusiform, sometimes flattened; pectoral and pelvic fins usually present and not encased in armour; jaws (when well preserved) of tusklike dermal bony elements. Some early groups freshwater; later groups with giant species (up to 10 m [roughly 33 ft]), marine.

†*Order Rhenanidiformes.* Found throughout the Devonian, but more common in first half of the period. Covered with small bony plates, head shield present in some; body flattened, raylike, with eyes on top of head; gill chambers typically placoderm in occupying a large area below head; pectoral fins greatly enlarged; transverse jaws armed with teeth; apparently primarily marine. Average length about 24 cm (roughly 9½ in.).

†*Order Antiarchiformes* (antiarchs). First known from Middle Devonian, extinct by end of the period. Head and thorax shield present; internal skeleton partially ossified in some; body fusiform but flattened ventrally for bottom living; pectoral fins movable but encased in armour; jaws of small transversely placed bony plates; eyes close together on top of head; well-preserved specimens show intestine with a spiral valve and lunglike structures; apparently mostly small bottom-dwelling freshwater fishes. Length about 10–40 cm (roughly 4–16 in.).

†*Order Stensioelliformes.* Lower Devonian. Not well-known but appearing to have a general lack of bone development, isolated tubercles covering skin in some; gill bars and jaws well developed in reasonably well-preserved specimens; marine. Small; length about 25 cm (roughly 10 in.).

†*Order Palaeospondyliformes.* Middle Devonian. Relationships uncertain, not a typical placoderm. Dermal armour lacking; ring-shaped vertebral centra and neural arches present; jaws apparently present. One genus, *Palaeospondylus,* many specimens, probably of a single species. Length about 4 cm (roughly 1½ in.).

CLASS SELACHII, or CHONDRICHTHYES (sharks, skates, rays, and relatives)

Vertebrates with jaws; dermal and endochondral bone absent, cartilage often calcified but no true bone except possibly at base of teeth and denticles (controversial); scales placoid, of dentine and enamel, present over entire body and enlarged to form teeth on jaws; scales and teeth do not grow once fully formed but are replaced when worn out; notochord often reduced, partially replaced by cartilage, which joins the connective tissue covering of the notochord; labial cartilages present in some; spiracle present (sometimes lost); claspers (pterygopodia) often present in pelvic fins of males, used in mating; intestinal spiral valve present in modern forms (condition in fossils unknown); lungs or structures similar to swim bladders are absent in modern forms.

†*Order Cladoselachiformes.* Middle Devonian to close of Permian. Notochord persistent in adult; 2 dorsal fins; each with a spine; no anal fin; basal cartilages or cartilages of pectoral fin remain along base of fin, radial cartilages unjointed;

pelvic fins without claspers; tail fin externally almost symmetrical but internally heterocercal; jaws amphistylic (upper jaw articulates with cranium and hyoid bone); rostral region of cranium small; postorbital process of cranium large; teeth with 1 large median cusp and smaller lateral cusps; 5 gill openings; marine predators.

†*Order Cladodontiformes.* Middle Devonian to about end of Permian. Notochord persistent in adult; 2 dorsal fins each with or without dorsal spine; anal fin absent; basal cartilage or cartilages of pectoral fin remain along base of fin; radial cartilages unjointed; pelvic fins with claspers; tail fin externally equilobate but internally heterocercal; jaws amphistylic; rostal region of cranium small; postorbital process of cranium large; teeth with 1 large median cusp and smaller lateral cusps; 5 gill openings; marine predators.

†*Order Xenacanthiformes.* Upper Devonian to Middle Triassic. Notochord persistent, but reduced, in adult; some cartilage present; 1 long-based dorsal fin, no spines; postoccipital head spine present; 2 structures similar to anal fins present; basal cartilages of pectoral fins occur in series and enter fin as axial element, radial cartilages unjointed; pelvic fins with claspers; tail fin diphycercal; rostral region of cranium small and postorbital process of cranium large; teeth with a small central cusp and a large lateral cusp on each side; 5 gill openings; freshwater predators.

Order Heterodontiformes. Upper Devonian to Recent. Notochord persistent in primitive forms, replaced partially in advanced ones; 2 dorsal fins, each with a spine; anal fin present; 3 basal cartilages in pectoral fin or a modification of this condition; fin more mobile than 3 preceding orders, radial cartilages of pectoral fins jointed; pelvics with claspers; tail heterocercal or modified from this; jaws remain amphistylic but trend toward hyostylic (supported by movable hyomandibular cartilage); postorbital process of cranium large to reduced; rostral region of cranium usually small; teeth cladodont-like to very modified and rounded; 5 gill openings; marine, several modern forms mollusk eating in habit (heterodonts). Includes the more primitive fossil hybodonts and more specialized, living, heterodonts or hornsharks (Heterodontidae).

Order Hexanchiformes. Jurassic to Recent. Notochord persistent but in some constricted anteriorly; 1 posterior spineless dorsal fin present; anal fin present; basal cartilages of pectoral fin reduced in number; radial cartilages of pectoral fins jointed; pelvic fins with claspers; tail heterocercal to nearly diphycercal; jaws essentially amphystylic but contact with hyoid arch absent in Hexanchidae; postorbital process of cranium reduced; teeth trifid in Chlamydoselachidae, many-pointed in Hexanchidae; 6 to 7 gill openings; marine predators; 2 living families, Chlamydoselachidae (frilled shark) and Hexanchidae (cowsharks).

Order Lamniformes (typical sharks). Lower Jurassic to Recent. Notochord in adults replaced by calcified cartilaginous centra; 2 dorsal fins present, opened or not; anal fin present or absent; 2 basal cartilages of pectoral fin reduced (except Orectolobidae, which has 2); pelvics with claspers, and with basal cartilage; tail heterocercal; jaws hyostylic, mobile, shortened and protrusible; postorbital process of cranium reduced; teeth with varied cusps; 5 gill openings; rostral area elongate; about 15 living families, typical sharks, mostly marine predators, free-swimming and bottom-dwelling, few in freshwater. Length to about 20 m (roughly 66 ft).

Order Pristiophoriformes (sawsharks). Cretaceous to Recent. Like Lamniformes, but with 6 gill openings. Elongated, flattened snout with sawlike teeth along sides in Recent members; body somewhat flattened but elongated; marine shore fishes and in fresh water, tropics. Length (in modern species) to about 1.2 m (roughly 4 ft).

Order Rajiformes (rays, banjofishes, and sawfishes). Upper Jurassic to Recent. Notochord replaced with calcified cartilage; 2, 1, or no dorsal fins; spines absent or present; anal fin absent; pelvic fins with claspers; tail heterocercal to modified, whiplike; jaws modified, supported by pseudohyoid cartilage, very mobile; 5 gill openings; rostral area elongate; marine bottom-dwelling sharklike fishes, flattened; spiracle (lateral opening) used for intake of water to gill chamber; eyes on top of head; gills ventral; greatly enlarged pectoral fins extend forward along gill opening, attached to sides of head and even meet in front of head in some; swim by wavelike motions of pectoral fins; 8 extant families, including Torpedinidae (electric rays). Medium to large fishes; maximum width (in manta ray) to 7 m (roughly 23 ft), weight to 1,700 kg (roughly 3,750 lb); length (in sawfish) to 11 m (roughly 36 ft) with weight to at least 2,500 kg (roughly 5,500 lb).

CLASS HOLOCEPHALI

Jaws holostylic (the palatoquadrate) supporting the upper jaw completely fused to cranium; hyoid arch complete, unmodified; branchial arches below cranium; internal skeleton of cartilage, often calcified but never of bone; dermal skeleton of dentine or dentine-like tissue (placoid scales), never with true bone; scales do not continue to grow once fully formed; pelvic and cephalic claspers in males of some groups.

Order Chimaeriformes (chimaeras). Upper Devonian to Recent. Teeth in a single series of a few tooth plates along each jaw ramus (half); pectoral with 2, and pelvic fins with 1 basal element; pelvic fin claspers present; dermal armour frequently present on head; primitive forms with placoid scales covering body, lost in certain advanced forms; scales specialized in some; dorsal fin spine present or absent; cephalic clasper present in some; marine.

†*Order Copodontiformes.* Devonian to Carboniferous. Known from teeth only; relationship uncertain. Marine.

†*Order Psammodontiformes.* Lower to Upper Carboniferous. Teeth only; little known. Marine.

†*Order Helodontiformes.* Lower Carboniferous to Upper Permian. Teeth numerous, about 10 series on each jaw ramus, some fused into tooth plates; no specialized symphyseal teeth; no cephalic clasper; pelvic claspers unknown; placoid scales cover body; marine.

†*Order Petalodontiformes.* Fossil only; Lower Carboniferous to Upper Permian. Known from teeth only, relationships uncertain; marine; Europe, Asia, North America.

†*Order Edestiformes.* Fossil only; Lower Carboniferous. Known only from specialized symphyseal (fused) teeth; marine; Europe, North America.

†*Order Chondrenchelyiformes.* Lower Carboniferous. Upper jaw with 4 pairs of tooth plates; lower jaw with 3 pairs of tooth plates; dermal plates on skull, cephalic clasper and dorsal fin spine absent, dorsal fin long, continuous along back; marine; known only from Scotland.

CLASS SARCOPTERYGII (fleshy-finned fishes)

Primitive members of the following 2 orders show certain similarities and so are placed together in this class. Some of these similarities are: Heterocercal tail fin with a small amount of fin development above the vertebral column at the posterior end of the tail fin; 2 dorsal fins present; pectoral fins an archipterygium of variable form (an axial median support with side branches); cosmoid scales present, similar to that in acanthodians; modern freshwater forms have lungs; presumably lungs were present in fossil freshwater forms also. Scales grow throughout life of the individual. The internal nares of the Crossopterygii and the Dipnoi may or may not have the same origin.

Order Crossopterygii (coelacanths and fossil relatives). Lower Devonian to Recent. Cranium divided into 2 parts (anterior and posterior) at region for exit of the 5th cranial nerve, these parts movable on each other; choanae (internal nares) present (lost in coelacanths); teeth labyrinthodont (*i.e.*, with complicated unfoldings of the enamel surface); 2 important groups, suborder Rhipidistia, Lower Devonian to Early Permian, mostly shallow freshwater and thought to have given rise to terrestrial vertebrates during the Devonian, and the suborder Coelacanthini. Upper Devonian to Recent, mostly marine, includes the so-called living fossil, *Latimeria chalumnae,* from South Africa, which lacks lungs. Length of rhipidistians to about 3 m (roughly 10 ft); of coelacanths to about 2 m (roughly 6½ ft).

Order Dipnoi (lungfishes). Lower Devonian to Recent. Cranium not divided into movable parts; teeth on upper jaw early reduced and lost in later members; pterygoid bones with fused teeth in plates modified for eating mollusks; 3 surviving types of lungfishes, 1 each in Australia, Africa, and South America. Length 60–200 cm (roughly 24–80 in.).

CLASS ACTINOPTERYGII (ray-finned fishes)

Fins supported by rays of dermal bone rather than by cartilage or cartilage bones. A group of jawed fishes so diverse that no single definition for them can be derived; better understood by determining the distinctive characters of the primitive members and then tracing their various lines of evolution. Primitive actinopterygians can be separated from the sarcopterygians by the following characteristics. Scales ganoid; single dorsal fin; pectoral fins with a series of thin radial bones, rather than basal plates and fleshy lobes; no internal nares. Other important characters: skeleton usually well ossified; scales grow throughout life; swim bladder present (occasionally modified to a lunglike structure).

Infraclass Chondrostei

A mixed group that has undergone many evolutionary diversifications. The remaining orders of the Chondrostei are specialized, often for special habitats and ways of life, but many of the groups show trends toward the holostean level of organization, especially in median fin structure and the development of hemiheterocercal tail in which externally at least the tail appears nearly homocercal.

†*Order Palaeonisciformes.* Lower Devonian to Middle Cretaceous. Mostly fusiform fishes with heterocercal tail; maxillary bone of the upper jaw bound to the preopercle bone space for the muscle restricted to lower jaw, limiting its power and func-

tion; many more fin rays than basal elements in the median fins; 37 families of wide distribution, early members freshwater, later marine.

†*Order Tarrasiiformes.* Carboniferous. Palaeoniscid-like, but with elongate body, a diphycercal tail and dorsal and anal fins continuous with it. One family, Tarrasiidae, Scotland and Illinois.

†*Order Haplolepiformes.* Upper Carboniferous. Peculiar fishes with stout unbranched fin rays; large gular plates; small opercular apparatus. One family, Teleopterinidae; Europe and North America.

†*Order Perleidiformes.* Lower to Upper Triassic. With ganoid scales; fin rays equal number of basal supports rather than exceed them as in Palaeonisciformes and other Chondrostei; tail hemiheterocercal. Three families; worldwide.

†*Order Redfieldiiformes.* Lower and Middle Triassic. Like Perleidiformes but fin rays more numerous than basal elements in dorsal and anal fins. One family, Dictyopygidae, fresh water of South Africa, Australia, and North America.

†*Order Dorypteriformes.* Upper Permian. Similar to Bobasatraniiformes but with very modified skull; scales confined to anterior part of trunk. One family, Dorypteridae; Europe, China.

†*Order Bobastraniiformes.* Lower Triassic. Body deep, laterally compressed; fin rays slightly more numerous than basal supports; opercular apparatus with small opercle, large preopercle; crushing dentition; pelvics absent. Thought to perhaps have been a coral feeder. One family, Bobasatraniidae; marine; widely distributed.

†*Order Pholidopleuriformes.* Lower to Upper Triassic. Some relatively long and slender; dorsal and anal fins far back on body, origin of anal fin anterior to dorsal fin; fin rays more numerous than basal elements; tail hemiheterocercal; jaw support almost vertical or moderately oblique, rather than extremely oblique as in most Chondrostei. One family, Pholidopleuridae; marine and freshwater; wide distribution.

†*Order Peltopleuriformes.* Upper Triassic. Large eyes; hemiheterocercal tail almost symmetrical externally; dentition weak. Two families, Peltopleuridae and Habroichthyidae; marine, perhaps some plankton feeding; Italy, China.

†*Order Platysiagiformes.* Lower Triassic to Lower Jurassic. Elongate, fusiform body, tail hemiheterocercal; median fins holostean, in that rays probably equalled basal elements; teeth large, conical. One family, Platysiagidae; marine; probably predacious; Italy and England.

†*Order Cephaloxeniformes.* Middle to Upper Triassic. Body deep, fusiform; thick head bones and crushing dentition; tail hemiheterocercal. One family, Cephaloxenidae; marine; probably bottom-dwelling mollusk eaters; Italy.

†*Order Luganoiiformes.* Middle and Upper Triassic. Almost holostean in character; body fusiform; head somewhat flattened in the horizontal plane; some head bones fused; jaw suspension inclined forward; fin rays apparently equal to basal elements in number; tail hemiheterocercal. One family, Luganoiidae; marine; probably predacious midwater fishes; Italy.

†*Order Ptycholepiformes.* Middle Triassic to Upper Jurassic. Structure near that of holosteans; fusiform body; fin rays of median fins nearly equalling basal elements in number; jaw support almost vertical; teeth small. One family, Ptycholepididae; marine; presumably plankton feeders; Europe.

†*Order Saurichthyiformes.* Lower Triassic to Upper Jurassic. Elongate, slender; snout elongate; single dorsal fin far back on body, opposite anal fin; tail diphycercal in appearance; number of scale rows reduced, 1 dorsal, 1 ventral, and 1 along each side; jaw suspension almost vertical; teeth large, conical, jaws long. One family, Saurichthyidae; marine and freshwater; predacious; worldwide. Length about 7–150 cm (roughly 2¾ to 60 in.).

†*Order Chondrosteiformes.* Lower Triassic to Upper Jurassic. Body scales and skull bones reduced; snout moderately developed, maxillary and opercular bones reduced; jaw support somewhat inclined backward; median fins paleoniscid-like, rays more numerous than basal supports. Probably gave rise to sturgeons. One family, Chondrosteidae; marine; some were suctorial feeders like sturgeons; England.

†*Order Parasemionotiformes.* Lower Triassic. Very near holosteans in structure but preopercle large and true suborbital bones still present, as in chondrosteans. Two families; marine; Siberia, Greenland, and Madagascar.

Order Acipenseriformes (sturgeons and paddlefishes). Upper Cretaceous to Recent. Almost no internal ossification; scales as large scutes in isolated rows (Acipenseridae); snout enlarged and tactile (Polyodontidae); median fins chondrostean in having more fin rays than basal elements; tail heterocercal. Marine and freshwater, bottom suctorial feeders (sturgeons, Acipenseridae; Europe, Asia, North America) and plankton feeders (paddlefishes, Polyodontidae; China and North America). Length (sturgeons) up to 9 m (roughly 30 ft), weight to 1,400 kg (roughly 3,100 lb).

Order Polypteriformes (bichirs and reedfish). Pleistocene to Recent. Relationships controversial, placed in own subclass by some and thought related to crossopterygians by others. Typical chondrostean characters, such as ganoid scales and a paleoniscoid type of preopercle. Fins modified into long continuous dorsal, tail diphycercal; freshwater; Africa.

Infraclass Holostei

Tail hemiheterocercal; maxillary scale free of preopercle; rays of median fins about equal basal elements in number; spiracle lost; vertebral column tended to increasing ossification; trend toward thinning scales and loss of ganoid layer.

Division Holosteans

Preoperculum intimately bound to and supporting the posterior border of the palate.

Order Amiiformes (bowfin and fossil relatives). Upper Triassic to Recent. Relatively conservative holosteans with typical holostean characters as given above; some specialized in body shape (elongate); most typical fusiform holosteans. One living member of the family Amiidae, with 1 species, *Amia calva* (bowfin), of North America; marine and freshwater, almost worldwide; 6 families.

†*Order Pachycormiformes.* Lower Jurassic to Upper Cretaceous. Long snout, suggesting the teleost swordfishes, Xiphiidae. Two families; Europe and North America.

Order Semionotiformes (gar pikes and fossil relatives). Upper Permian to Recent. Two families of widely divergent fishes; probably independent of the Amiiformes but with typical holostean characters; the fossil Lepidotidae with normal holostean fusiform bodies, which become relatively deep and slab-sided in some members; marine and freshwater, widely distributed. Gar pikes (Lepisosteidae) are elongated, sharp snouted, primarily freshwater predators, still extant in North America; length to about 3.5 m (roughly 11½ ft).

†*Order Pycnodontiformes.* Lower Jurassic to at least Eocene. Very deep bodied, with jaws and teeth modified for nibbling; perhaps fed on coral; marine and widespread.

†*Division Halecostomes*

Holosteans but with a preoperculum not buttressing the bones of the palate.

†*Order Pholidophoriformes.* Upper Triassic to Upper Cretaceous. Difficult to separate from the more primitive of the teleost orders (below). Holosteans with some trends toward teleosts, notably: loss of ganoine from fin rays, scales, and dermal bones; loss of peg and socket joints between scales; loss of bone cells in scales retained in some teleosts; development of intermuscular bones; loss of scalelike bones (fulcra) on leading edges of fins (retained in caudal fin of some teleosts); loss of some skull bones; a "sinking" of skull bones beneath the skin. The caudal skeleton is the major difference between the pholidophoroids and teleosts. Pholidophoroids have the caudal centra incompletely ossified and lack "splint" bones called uroneurals (modified neural arches) that give ridged support to the terminal 4 or 5 vertebrae in teleosts. About 7 families, of which the Jurassic Pholidophoridae are the most likely ancestors of the teleosts; marine and freshwater, of wide distribution.

Infraclass Teleostei (bony fishes)

Tail homocercal; caudal skeleton with perichordally (around the spinal chord) ossified centra; neural arches modified into elongate uroneurals extending forward onto the preural centra, "stiffening" the joints between the terminal 4 or 5 vertebrae. Two hypural bones supporting the lower caudal fin lobe. (Note: Although the above statement can be used to define and separate early teleosts from holosteans, many later groups of teleosts have modified the tail structure greatly, so the definition will not "fit" at first sight. It can be readily shown, however, that all teleosts have a caudal structure derived from that described above.) Teleosts never have ganoid scales; typically, their scales when present are thin, overlapping plates of bone that continue to grow throughout life; their lower jaws lack certain bones found in many chondrosteans or at least have some of these bones fused to single elements.

†*Superorder Leptolepidimorpha.* The recognition of this superorder is highly tentative pending determination of the relationships of these fishes with other teleosts. Sometimes classified with the Halecostomi, these fishes were clearly teleosts in their caudal fin structure. Preopercle supported the palate (as in some holosteans) but with several shifts in this region to a teleost-like preopercular-jaw arrangement and with the adductor muscle of the mandible attached to preopercle; no bone cells in scales; more than 1 supraorbital; gular plate present; apparently no adipose fin; rostral elements with a bone enclosed commissure.

†*Order Leptolepidiformes.* Triassic to Middle or Upper Cretaceous. The characters of the order are those listed above for the superorder. Four families; widely distributed.

Superorder Elopomorpha. A diverse group including very primitive fishes and specialized fishes such as eels and therefore difficult to define. Some primitive members with a gular plate (absent in eels), ethmoid commissure present in some forms in a dermal rostral bone (lost in many eels); a leptocephalus larva; no bone cells in scales of primitive members; pelvic fins abdominal when present.

Order Elopiformes (tarpons, tenpounders, and bonefishes). Upper Jurassic to Recent. Body fusiform, typical fishlike shape; bone-enclosed ethmoid commissure present; roofed post-temporal fossae; primary bite a tongue-parasphenoid type; marine; worldwide in temperate and tropical zones.

Order Anguilliformes (eels). Cretaceous to Recent. Body elongate; fins reduced and gill chamber modified; displaced posterior to much of head; opercular apparatus reduced; pectoral girdle free of skull; caudal and other fins often greatly reduced; bony ethmoid commissure sometimes present; marine and freshwater, worldwide in temperate and tropical regions. Length about 15–300 cm (roughly 6 to 120 in.).

Order Notacanthiformes (deep-sea spiny eels). Middle Cretaceous to Recent. Ethmoid commissure present, like that of elopomorphs; body relatively elongate and tail skeleton reduced; opercular apparatus complete. Three marine, deep-sea families, Halosauridae, Lipogenyidae, and the Notocanthidae (deep-sea spiny eels). Average length about 50 cm (roughly 20 in.).

Superorder Clupeomorpha. Special type of ear–swim-bladder connection present, consisting of a diverticulum of the swim bladder, forming bulla (cavity) within the ear capsule; head lateral line canals on operculum. A diverse group of mostly oceanic, silvery, compressed fishes, many of great commercial importance.

Order Clupeiformes (herrings, anchovies, and allies). Lower Cretaceous to Recent. Characters of the superorder; marine and freshwater, some anadromous; worldwide.

Superorder Osteoglossomorpha. A diverse group of freshwater fishes with a relatively primitive jaw suspension and shoulder girdle. The primary bite of the mouth between parasphenoid and tongue (basihyal and glossohyal); paired rods present, usually bony, at the base of the second gill arch; no bony ethmoid commissure; no leptocephalus larvae.

Order Osteoglossiformes (bony tongues, freshwater butterfly fishes, mooneyes, knife fishes). Middle Cretaceous to Recent. Circumorbital bones well-developed; scales with an irregular reticulated pattern (except Pantodontidae); freshwater, almost worldwide except extremely cold regions. Four families.

Order Mormyriformes (mormyrs). Pleistocene to Recent. With electricity-producing organs; orbital bones reduced; the cerebellum greatly enlarged; swim-bladder–ear connection reduced in adult; sometimes with long snouts for feeding in mud; 2 families, the elephant fishes, Mormyridae, and the Gymnarchidae; fresh water, Africa.

Superorder Protacanthopterygii. A diverse group of relatively primitive teleosts, mostly related by their lack of the specializations (or, in some cases, primitive features) found in the other teleost superorders. Vertebrae usually more than 24; adipose fin present in many members; mesocoracoid bone usually present; glossohyal teeth usually prominent (lost in some); upper jaw usually not protrusible; proethmoid and a series of several perichondral ethmoid commissures; 1 supraorbital bone; no gular plate.

Order Salmoniformes (salmons, trouts, whitefishes, smelts, pikes, and allies). Cretaceous to Recent. Characters are those given for superorder; endochondral ossification often somewhat reduced; marine and freshwater, worldwide. A large and important order, comprising about 35–40 extant and about 6 fossil families. Length about 4–115 cm (roughly 1½ to 45 in.); weight to about 50 kg (roughly 110 lb).

Order Ctenothrissiformes. Mostly Upper Cretaceous, marine fishes of uncertain affinities; possibly close to the basal stock from which the acanthopterygians are derived. The living marine family Macristiidae may belong here.

Order Gonorynchiformes (milkfish and certain deep-sea fishes). Cretaceous to Recent. Toothless; with epibranchial organs and a characteristic caudal skeleton; marine of Indo-Pacific and freshwater of Africa. The anterior ribs and vertebrae show affinities with the superorder Ostariophysi, and the group may belong with the ostariophysans rather than with the Protacanthopterygii. Length about 10–150 cm (roughly 4 to 60 in.).

Superorder Ostariophysi. A group of 5,000–6,000 species, including the majority of known freshwater fishes. Characterized by possession of a Weberian apparatus (a swim-bladder–internal-ear connection with three movable bones).

Order Cypriniformes (characins, tetras, some knife fishes, carps, and minnows). Lower Eocene to Recent. Parietal, symplectic, and subopercular bones present; worldwide in fresh water except Antarctica and Australia. A few North Asian forms enter the sea.

Order Siluriformes (catfishes). Paleocene to Recent. Parietal, symplectic, suboperculum, and true scales absent; often with dermal plates or little bony spines in the skin. Fusion of the supportive parts of the Weberian apparatus extensive. About 30 families; distribution of the superorder primarily freshwater but some families marine, with the majority of the 2,000 species in Africa and South America.

Superorder Scopelomorpha. This superorder, like the Paracanthopterygii and Acanthopterygii (below), is characterized by a tendency toward fin spines, subocular shelves, and separated exoccipital condyles. Scopelomorphs frequently retain such primitive structures as an adipose fin and an asymmetrical caudal fin skeleton.

Order Myctophiformes. Cretaceous to Recent. Characteristics of the superorder. Distinctive jaw musculature, like that of Paracanthopterygii (below). Two subgroups of this order differ in ecology and structure. One contains several families of deep-sea fishes, often elongated predators with large teeth. The second contains benthic fishes, or bottom dwellers (*e.g.*, Aulopidae), tropical inshore fishes (Synodontidae, or lizardfishes), midwater deep-sea fishes with light organs (the large family Myctophidae, or lantern fishes), and bottom-dwelling deep-sea fishes (*e.g.*, Bathypteroidae, or spiderfishes). Order contains about 15 families, worldwide; marine. Mostly small fishes 10–15 cm (roughly 4 to 6 in.); maximum length about 95 cm (roughly 37½ in.).

Superorder Paracanthopterygii. Most with a distinctive type of jaw musculature (involving levitor maxillae superioris muscle and associated structures); caudal vertebrae with the 2nd ural centrum fused with the upper hypural, 2 or fewer epurals and a full neural spine on the 2nd preural centrum; pelvic fins usually placed anteriorly, thoracic (midbody) or even farther forward. In general these fishes have tended to lose primitive acanthopterygian characters.

Order Polymixiiformes (barbudos). Middle Cretaceous to Recent. Barbels suspended from the hypohyal bones (anterior part of the gill arches); spines on the dorsal and anal fins; pelvic fins subthoracic. Retain some primitive paracanthopterygian characters, such as an antorbital bone, a free 2nd ural centrum, 6 autogenous hypurals, 2 uroneurals, and Baudelot's ligament to the 1st vertebra. Adipose fin lacking. Deepwater marine fishes; 1 family, probably only 2 species. Adult length about 30 cm (roughly 12 in.).

Order Percopsiformes (trout-perches, pirate perches, and cave fishes). Eocene to Recent. Mouth gape and buccal dentition reduced; median fin spines reduced or lost; head with spine ornamentation; scale covering of the adipose fin lost. All living species freshwater, North America; length 8–13 cm (roughly 3 to 5 in.). Three extant families, 1 fossil family.

Order Gadiformes. Lower Eocene to Recent. Early gadiforms were similar in structure to early percopsiforms, but almost all remained marine and subsequently specialized into a variety of environments. Reduced caudal skeleton; elongate body; altered head and jaw structure. Primitive gadiforms have 7 branchiostegal rays, primitive percopsiforms 6. All with very reduced fin spines; marine, worldwide. Order includes cods, hakes, cusk eels, pearlfishes, eelpouts, grenadiers, and rattails. Length 7 to about 200 cm (2¾ to 79 in.).

Order Batrachoidiformes (toadfishes). Miocene to Recent. Bottom fishes with short, small, spinous dorsal fins; long soft-rayed dorsal fins; flat heads; 1 family, Batrachoididae; marine, occasionally freshwater, shore fishes of tropics. Length to about 40 cm (15¾ in.).

Order Lophiiformes (goosefishes, anglerfishes, frogfishes and batfishes). Eocene to Recent. Spinous dorsal fin modified as a movable lure. Some deep-sea forms with light organs and males parasitic on females. Marine, widespread; in shallow-water and deep-sea habitats. About 15 families. Length to about 130 cm (51 in.).

Order Gobiesociformes (clingfishes). Recent questionable fossil from Miocene of California. Flattened, depressed fishes with a ventral sucker formed of the pelvic fin and surrounding tissue; no spiny dorsal fin; 1 family, Gobiesocidae; marine and occasionally freshwater in tropics and along many temperate seacoasts.

Superorder Acanthopterygii (spiny-rayed fishes). Spiny fins usually emphasized, rather than reduced (as in paracanthopterygians). Mobile, protractile mouth due to the almost universal lack of the levator mandibula superioris muscle; pectoral fin relatively higher on side of body; Baudelot's ligament almost always attached to basicranium. The 13 orders of the superorder Actinopterygii may be divided into 2 categories (sometimes called series) on the basis of the number of vertebrae, the condition of the fin spines, the position of the pelvic fins, and the presence or absence of ctenoid scales. The series Atherinomopha contains only the order Atheriniformes, the series Percomorpha the remaining actinopterygian orders.

Order Atheriniformes. Eocene to Recent. Fin spines present or absent but frequently weak when present; vertebral number higher than 24; ctenoid scales rare; pelvic fins abdominal, subabdominal, or thoracic in position. Marine shore fishes, also freshwater, tropical and temperate, worldwide. About 16 families, including the oceanic flying fishes (Exocoetidae), killifishes (Cyprinodontidae), live-bearing topminnows (Poeciliidae), and silversides (Atherinidae). Mostly small fishes (2–10 cm [roughly ¾ to 4 in.]) but some needlefishes (Belonidae) to about 130 cm (51 in.).

Order Lampridiformes. Paleocene to Recent. Intermediate in some ways between polymixioids and acanthopterygians, but all living members very specialized. All lack a subocular shelf and pelvic spine; some have a peculiar condition (hypurostegy) in which caudal rays are expanded. Marine, oceanic, tropic, and temperate regions. About 9 families. Medium to large size; to about 2 m (6½ ft) and 300 kg (660 lb) in the opah (Lampridae) and about 10 m (32¾ ft), but far less weight, in the more slender oarfishes, Regalecidae (treated below with the flying fishes and others in the order Atheriniformes).

Order Beryciformes. Squirrelfishes and several deep-sea fishes. Cretaceous to Recent. Spines present in fins; pelvic fins subthoracic; retained primitive number of caudal branched fin rays (17), primitive number of epurals (3); exoccipital condyles poorly developed; orbitosphenoid present. About 15 families of small to medium-sized fishes. Length 5–60 cm (roughly 2 to 23½ in.). Marine, worldwide in tropical and temperate regions (treated below with the flying fishes and others in the order Atheriniformes).

Order Zeiformes (John Dories, boarfishes, and relatives). Lower Eocene to Recent. Anal fin with 1–4 spines; pelvic fin with 1 spine and 5–9 branched rays; caudal fin with less than 15 principal rays; about 6 families, of which the dories, Zeidae, are best known; marine, deep-sea, widespread. Length to about 1 m (3¼ ft) (treated below with the flying fishes and others in the order Atheriniformes).

Order Gasterosteiformes (sticklebacks, tube-snout fish, and sea horses). Eocene to Recent. Frequently with strong spines in dorsal and pelvic fins, spines absent in some; snout often elongated; body often with dermal plates; 9 families, marine and freshwater, widely distributed. Length about 3–200 cm (1¼ to 78¾ in.).

Order Channiformes (snakeheads). Pliocene to Recent. Elongate bodies; dorsal and anal fins present; depressed head with an accessory air-breathing apparatus in the gill chamber. Snakeheads, Channidae; freshwater, tropical Old World. Length about 15–95 cm (6 to 37½ in.).

Order Synbranchiformes (swamp eels). No fossil record. Fins reduced, fin spines absent, pharynx modified for breathing air. Three families, restricted to freshwater, in tropics. Length 20 to about 50 cm (roughly 8 to 20 in.).

Order Scorpaeniformes (scorpionfishes, sculpins, and relatives). Eocene to Recent. A complex group of widely divergent fishes that may be polyphyletic and is difficult to characterize. Three groups may be recognized: scorpaenoid, hexagramoid-cottoid, and anoplopomatoid. United as an order because of a distinctive caudal skeleton and a bony process connecting the 3rd orbital with the preoperculum in most members. Some members with external bony plates. About 21 families; primarily marine, some freshwater, in tropical and temperate regions.

Order Dactylopteriformes (flying gurnards). Pliocene to Recent. Bottom-dwelling shore fishes with dermal armoured plates, movable modified pectoral fin rays. Marine. One family, Dactylopteridae.

Order Pegasiformes (dragonfishes). No fossil record. Bottom-dwelling marine fishes with dermal armour and large pectoral fins. One family, Pegasidae.

Order Perciformes. Upper Cretaceous to Recent. Fins usually with spines; pelvic fin with 1 spine and not more than 5 rays, usually below pectoral fins; caudal fin with 15 rays; no orbitosphenoid, mesocoracoid, or intermuscular bones. An extremely varied assemblage of fishes, with a variety of body plans and other adaptations. About 140 families, divided among about 20 suborders. Mostly marine, worldwide. Size shows broad range; adult length from about 1 cm (less than ½ in.) (certain gobies) to about 4.8 m (15¾ ft) (swordfish); weight to about 900 kg (roughly 2000 lb).

Order Pleuronectiformes (flatfishes). Eocene to Recent. Both eyes on same side of head, skull twisted and asymmetrical, fins usually without spines. Seven families include flounders, soles, and halibuts; mostly marine; bottom fishes, worldwide in tropical and temperate regions.

Order Tetraodontiformes. Eocene to Recent. With a beaklike snout, gill opening restricted to a small opening; probably related to acanthuroids. Eleven families; marine, occasionally freshwater, worldwide in tropics and subtropics. (S.H.W.)

THE PRIMITIVE FISHES (LIVING REPRESENTATIVES)

The jawless fishes: lampreys and hagfishes (Agnatha)

The class Agnatha (phylum Chordata) comprises the jawless, fishlike hagfishes and lampreys and several fossil groups. Hagfishes are minor pests of commercial food fisheries of the North Atlantic, but because of their parasitic habit, lampreys have been a serious pest of food fisheries in the Great Lakes in North America where they have reduced the numbers of lake trout and other species. Agnathans are otherwise of little economic importance. The group is of great evolutionary interest, however, because it includes the oldest known craniate fossils and because the living agnathans have many primitive characteristics.

GENERAL FEATURES

The body of the hagfish is softskinned and nearly cylindrical, with a single nostril at the anterior end, overlying the mouth, and a low caudal fin around the tail. The eyes are vestigial and covered by skin. All of the 20 known species are restricted to cold, marine bottom waters at depths ranging from 10 metres (about 33 feet) in high latitudes to 1,300 metres in equatorial oceans. Adults are 40 to 80 centimetres (15 to 30 inches) long. All species are superficially similar except in the number and position of the gill apertures.

Lampreys

Lampreys, which number about 22 species, are found in cool, fresh, and coastal waters of all continents except Africa. All species are rather similar. The body is smooth and eel-shaped, with well-developed dorsal and caudal fins; the mouth is surrounded by a suctorial oral disk bearing horny teeth. The eyes are well developed, and the nostril is on the top of the head. Adults range from 15 to 100 centimetres long.

Although the gonad of a hagfish usually includes both ovary and testis, there is no evidence either of hermaphroditism (the reproductive organs of both sexes functioning in the same individual) or of self-fertilization. A female produces a small number of tough-skinned, yolk-filled eggs about two centimetres long that hatch into miniature adults.

Hagfishes locate their food by scent. Although some are known to eat fishes immobilized in nets, the best studied species, *Myxine glutinosa,* normally feeds on soft-bodied invertebrates and larger dead animals. *Myxine* burrows into soft marine sediments and rests with only the tip of the head protruding. During respiration, water enters through the nostril and passes by a nasopharyngeal duct to the pharynx and gills. When stimulated by the scent of a dead fish, *Myxine* leaves its burrow and swims against the current. On making contact with the fish, it coils around it and bites into it by protruding and retracting the comblike horny tooth plates on the floor of the mouth. Along each side of the body is a row of prominent glands that produce a distasteful gelatinous slime when the animal is disturbed.

Lampreys mate in a nestlike depression excavated by the male in the gravel bed of a stream; the numerous eggs, about one millimetre (.04 inch) in diameter, lodge in the gravel around the nest. The egg hatches into an ammocoete larva—a blind, wormlike animal that burrows in silt. The larva's mouth is overhung by a hoodlike upper lip that protrudes above the surface of the silt. A continuous stream of water passes in through the mouth and out through the seven pairs of gills. Microscopic plants, the food of the ammocoete, are filtered from the respiratory current by strands of mucus produced by the endostyle, a gland in the floor of the pharynx.

After about three years, when the ammocoete has grown to about 10 centimetres, it undergoes a radical metamorphosis. The eyes complete their development; the upper lip becomes transformed into a suctorial oral disk; the endostyle changes into a thyroid gland; and the fins along the back increase in height. On completion of metamorphosis, a typical lamprey such as *Petromyzon marinus* migrates to the sea, where it feeds by attaching itself with its sucker to bony fishes. It rasps into the flesh with a toothed, tonguelike structure on the floor of the mouth. Saliva containing an anticoagulant facilitates the ingestion of blood and muscle tissue. On attaining full adult size, the lamprey ceases to feed, migrates upstream to a spawning ground, mates, and dies.

Members of several lamprey species do not migrate to sea but feed in fresh water. *Petromyzon marinus dorsatus* once seriously affected commercial fishing in the Great Lakes until measures were undertaken to control it. The brook lampreys do not feed after metamorphosis but mature sexually and reproduce.

From A. Brodal and R. Fange (ed.), *The Biology of Myxine*, Universitetsforlaget, Oslo

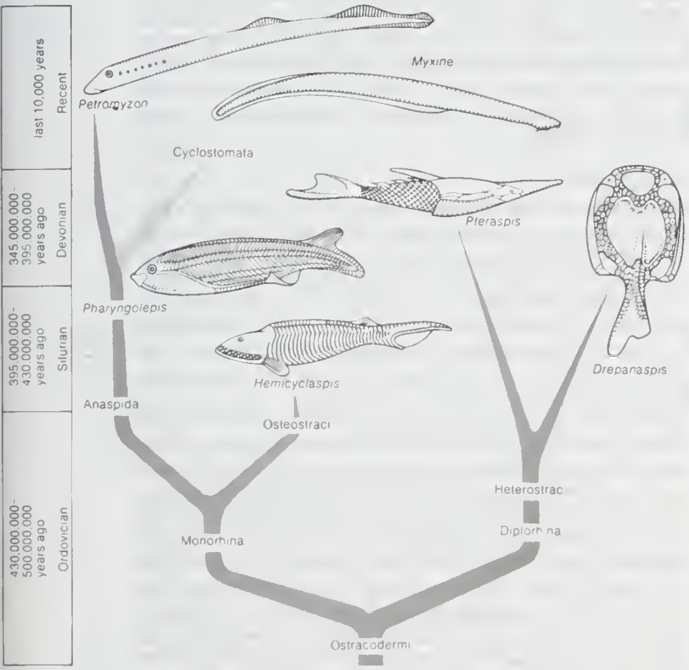


Figure 7: Systematic division of fossil and recent Agnatha.

FORM AND FUNCTION

Variation in body form

The elongated bodies of hagfishes and lampreys are adaptations to a burrowing habit—throughout life in the hagfish, during the larval period in the lamprey. Considerable variation in body form occurs among extinct agnathans. Some Osteostraci, for example, were flattened and, although possessing a powerful swimming tail, appear to have been bottom-dwellers. The size and shape of the mouth suggest that they were filter feeders. The laterally compressed, fishlike form of the anaspids (*e.g.*, *Pharyngolepis*) indicates a free-swimming habit; the absence of fossilized impressions of the mouth region suggests the presence of large suctorial or grasping lips. The extinct heterostracans include obvious bottom-dwellers (*e.g.*, *Drepanaspis*) and others (*e.g.*, *Pteraspis*) that were apparently adapted to mid-water, or nektonic, life. Some heterostracans, for example, had movable enamel plates inside the lower lip, probably to provide a biting or grazing mechanism. Osteostracans, anaspids, hagfishes, and lampreys (Monorhina) have one median nostril, but heterostracans (Diplorhina) appear to have had two, one at each corner of the mouth.

EVOLUTION AND CLASSIFICATION

Evolution. If evidence from fossil and living forms is combined, the Agnatha are distinguishable from the other craniates (Gnathostomata) by what they lack: jaws, lateral fins supported by fin rays, vertebrae, a horizontal semicircular canal in the ear, and genital ducts. The nervous, sensory, endocrine, circulatory, excretory, and muscular systems have the same basic structure as those of gnathostomes but are generally simpler. The presence in the ammocoete larva of an endostyle, a gland that otherwise is found only in protochordates (*e.g.*, amphioxus, tunicates), suggests that the Agnatha represent an evolutionary level intermediate between the protochordates and gnathostomes; but the degree of specialization of known agnathans (particularly the single nostril of most forms) rules out the possibility that they are ancestral to the gnathostomes.

Annotated classification. In the classification below, the groups indicated by a dagger (†) are extinct and known only from fossils.

CLASS AGNATHA

Craniate chordates, lacking jaws; Silurian Period (395,000,000–430,000,000 years ago) to Recent.

Subclass Monorhina

With 1 nostril.

†*Order Osteostraci*

Late Silurian to Late Devonian (345,000,000–395,000,000 years ago); head and gills enclosed in heavy bony shield; nostril between eyes; gill openings ventral; about 7 families.

†*Order Anaspida*

Late Silurian to Late Devonian; active swimmers with tail bent downward; nostril between eyes; gill openings lateral; about 5 families.

Order Cyclostomata (lampreys and hagfishes)

Scales absent, skeleton cartilaginous, horny teeth in mouth, body eel-shaped.

Suborder Myxini (slime eels and hagfishes). Recent; nostril at anterior tip; 5–15 pairs of gills; 2 rows of horny teeth; eyes vestigial.

Family Eptatretidae. Five to 15 gills opening separately to exterior; all oceans except North Atlantic; 2 genera, 13 species.

Family Myxinidae. Five to 7 gills opening by common apertures on each side; all oceans; 4 genera, 8 species.

Suborder Petromyzones (lampreys). Pennsylvanian (280,000,000–325,000,000 years ago) to Recent; dorsal nostril, 7 pairs of gills, horny teeth on oral disk; eyes large, 1 or 2 dorsal fins.

Family Petromyzonidae. Single horny toothplate above the mouth, bearing pointed or rounded teeth; 5 genera, 17 species; Eurasia and North America.

Family Mordaciidae. Two tricuspid teeth above mouth; 1 genus, 4 species; eastern Australia and western South America.

Family Geotriidae. Single horny toothplate above the mouth, bearing 4 spatulate teeth; 1 genus, 1 species; southern Australia, New Zealand, South America.

Subclass Diplorhina

With 2 nostrils.

†*Order Heterostraci*

Ordovician to Upper Devonian (345,000,000 to 500,000,000 years ago); usually heavily armoured; gills with common aperture on each side; length 5–30 cm (2 to 12 in.); about 12 families.

†*Order Coelolepida*

Upper Silurian and Lower Devonian (395,000,000 years ago); a little-known group of unknown affinities; length to about 10 cm; 1 family.

Critical appraisal. Although lampreys have many features in common with the Osteostraci and Anaspida, general disagreement prevails on the classification of the hagfishes. Their affinity with lampreys is questionable, and they should perhaps be removed from the Cyclostomata as a separate order. It has been suggested that the hagfishes are closely related to the heterostracans, but this view is not widely accepted. (Ro.S.)

The cartilaginous fishes: sharks, skates, rays (Chondrichthyes, or Selachii)

The sharks, together with their close relatives, the rays, belong to one of the two great groups of living fishes, the class Chondrichthyes, or Selachii. The latter name is also used for an order that includes only the sharks. Many structural, physiological, biochemical, and behavioral peculiarities make these fishes of particular interest to scientists. The dissection of a small shark is the biology student's introduction to vertebrate anatomy. These fishes are, in a sense, living fossils, for many of the living sharks and rays are assigned to the same genera as species that swam the Cretaceous seas over 100,000,000 years ago. Although by any reckoning a successful group, the modern chondrichthyeds number far fewer species than the more advanced bony fishes, or teleosts; 200 to 250 species of sharks and 300 to 340 species of rays are known.

The danger some sharks and stingrays present to humans makes these animals fascinating and, at the same time, abhorrent. Perhaps for this reason, they figure prominently in the folklore and art of many tropical peoples whose living depends on the sea. The danger from shark attack, while very real, is easily sensationalized, and quite frequently little attempt is made to distinguish between dangerous and harmless species.

GENERAL FEATURES

Problems of taxonomy. The name Selachii refers to a category of fishlike vertebrates, which are given a variety of treatments by ichthyologists. Some consider the Selachii to be a class or subclass comprising all the modern sharks and rays; others restrict the name to an order limited to the modern sharks and certain extinct ancestral forms. Under the latter system, the rays (including the sawfishes, guitarfishes, electric rays, skates, and stingrays) are ranked as a separate order, and the two orders are placed in a class or subclass.

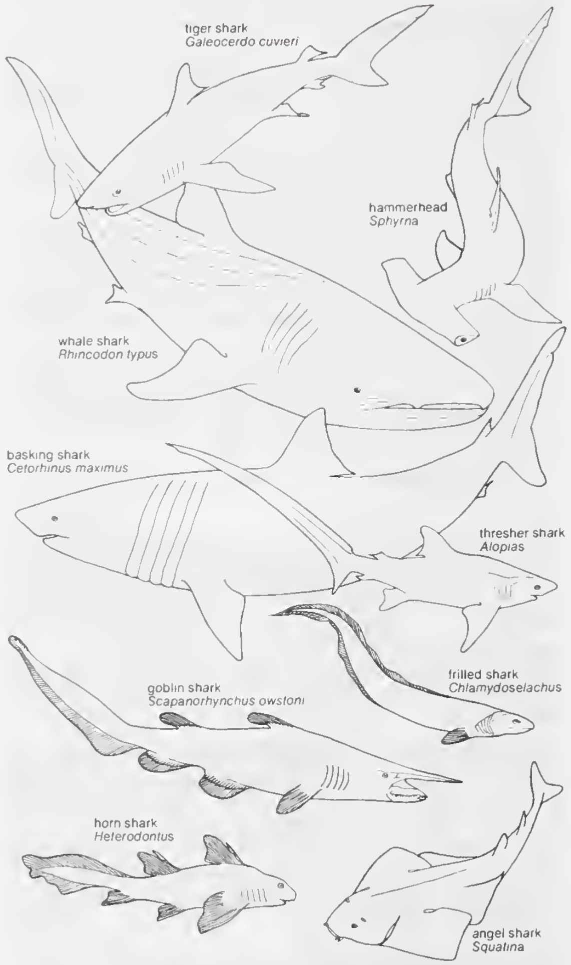


Figure 8: Body plans of representative sharks.

The chimaeras (Holocephali) bear many similarities to sharks and rays; *e.g.,* in skeletal structure, internal organs, and physiology. Ichthyologists commonly, although not unanimously, emphasize these similarities by grouping the modern and ancient sharks, rays, and chimaeras in the class Chondrichthyes, the cartilaginous fishes. Under this system, which is used in the present article, the sharks, skates, and rays are further grouped into one subclass, Elasmobranchii, and the chimaeras into another, Holocephali. (A classification in which the elasmobranchs constitute one class [Selachii] and the chimaeras another [Helocephali] is found above; see *Fishes: a comparative study: Annotated classification.*) Assigning the two groups class rank implies a degree of distinctness equal to that of the amphibians (Amphibia), reptiles (Reptilia), birds (Aves), and mammals (Mammalia).

Distribution and abundance. The majority of sharks and rays are marine fishes, but many enter estuaries; some travel far up rivers, and a few are reported to be permanent residents of freshwater. Most species live in the relatively shallow waters of continental margins or around offshore islands; a few roam far out in the vast spaces of the oceans. Some live at great depths, in midwaters or on the bottom; others are surface swimmers or inhabit the bottom in shallow waters.

Sharks and rays are poorly represented in fish markets of most countries. With limited demand for them, the damage they do to ordinary fishing gear, and the special care required to keep them marketable, fishermen avoid them if possible, or even discard those they happen to catch rather than bring them to port. Consequently, as a source of animal protein sharks and rays are generally underexploited, while the more highly valued bony fishes are generally overexploited. A possible consequence of this may be an increasing prominence of sharks and rays in the marine biota.

IMPORTANCE

Economic uses of elasmobranchs. *Sharks as food.* The meat of sharks is marketed for food in all maritime countries. It may be prepared in various ways—fresh, salted, smoked, or pickled—offered in such forms as steaks, fillets, or flakes and under such names as shark, whitefish, grayfish, swordfish, sea bass, and halibut. The flesh is often rather strong tasting; this quality, however, is one that can be removed by cleaning and washing and soaking the flesh in brine.

Since ancient times, Chinese people have used the dorsal fins of certain sharks and rays as the basis of an epicurean soup. To meet the demand for this product, they have imported fins from far-distant countries. The fins are prepared for market by removing the skin and flesh, leaving only the gelatin-rich cartilaginous rays, which are dried before shipment. Shark liver oil is used in various regions for tanning leather; for preserving wood; as a lubricant; as a folk medicine against rheumatism, burns, and coughs; as a general tonic; as a laxative; and as an ingredient of cosmetics. The liver of a basking shark yields 80 to 600 gallons of oil, which was used in lamps until petroleum products replaced animal oils for illumination. The discovery around 1940 that the liver of the soupfin shark of California is peculiarly rich in vitamin A led to an explosive development of a special fishery in California for this species and a search in other parts of the world for sharks having livers of comparable potency. Within a few years, however, the economic bubble burst, with the invention of a method for manufacturing synthetic vitamin A. The Australian school shark, which was used originally for vitamin A, is now caught for fish fillets.

Other shark products. The hard scales provide an abrasive surface to the skin of sharks and some rays, giving it a special value, as a leather called shagreen, for polishing hard wood. When heated and polished, shagreen is used for decorating ornaments and, in Japan, for covering sword hilts.

Shark leather is made in several countries, including the United States, from the skin of certain shark species after removal of the scales by a chemical process. A luxury product, much more durable than cowhide, shark leather is used for footwear, belts, wallets, and other accessories. The most suitable skins for leather are from tiger, dusky, brown, sand, blacktip, and nurse sharks.

In Greenland some Eskimos make rope from strips of the skin of the sleeper shark. Polynesians once added to the effectiveness of their war clubs with sharks' teeth. Sharks' teeth have some commerical value as curios. The Maori of New Zealand formerly paid high prices for mako sharks' teeth, which they wore as earrings.

Economic value of rays. About 126,000 tons (roughly 110,000,000 kilograms) of rays are marketed for food in various countries about the world, principally in Europe and Asia. By-products in local demand are skins of scaleless species for drumheads; those of scaly species are used for shagreen. Livers are used for oil, fins for gelatin. People of many tropical regions—Polynesia, Oceania, Malaysia, Central America, and Africa—have used the spines of stingrays for such items as needles and awls, spear tips and daggers, and for the poison they contain. The entire tails of stingrays, complete with spines, have been used as whips in various tropical areas.

The electric rays, or numbfish, have little commercial value. The ancient Greeks and Romans used the electric shock of *Torpedo* to relieve diseases of the spleen, chronic headaches, and gout. From the Greek word for electric ray, *narke,* comes the word narcotic. Today these fishes are of interest chiefly as a source of irritation (if not danger) to bathers who step on them and to fishermen who may be shocked when hauling in their wet nets. (L.A.Wa.)

Attacks by sharks

Danger to human life. Among the known shark species, 27 have been authoritatively implicated in attacks on persons or boats. Hospital and other records attest to many attacks on bathers, divers, and people awash in the sea following sea or air disasters. There are also many documented cases of sharks attacking small boats. A number of surviving victims have been able to identify the attacking animal as a shark; a few even reported the type of shark, such as a hammerhead. In many instances, witnesses have seen the assailant clearly enough to determine the species. Fragments of teeth left in wounds of victims or in the planking of boats have often been large enough to provide ichthyologists with the means for precisc identification.

In 1958 the American Institute of Biological Sciences established a Shark Research Panel at the Smithsonian Institution and Cornell University to gather historical and current records of shark attacks throughout the world. For the 35 years from 1928 to 1962, inclusive, the panel listed 670 attacks on persons and 102 on boats. Attacks occur most frequently throughout the year in the tropical zone between 21° north and south of the Equator; from midspring to midfall they extend as far north and south as the 42° parallels. For this reason, it was formerly believed that the most dangerous sharks lived in waters warmer than 21° C (70° F) and that the risk of attack was greatest in the tropics and in the summer months. It is now thought that this circumstance simply results from the fact that more people swim in warm water. It is known, for example, that the most dangerous shark, the white shark, or man-eater (*Carcharodon carcharias*), ranges into the cooler waters of both hemispheres. (L.A.Wa./Ed.)

In Australia, New Zealand, South Africa, and along other coasts heavily infested with sharks, public beaches have lookout towers, bells or sirens, and nets to protect bathers. Since 1937 Australia has used meshing offshore to catch the sharks. Gill nets suspended between buoys and anchors running parallel to the beach and beyond the breaker line have decreased the danger of attack. The nets enmesh sharks from any direction, and although they touch neither the surface nor the bottom, and are spaced well apart, they provide effective control. South Africa has used a similar protection system and has also conducted experiments with electrical barriers.

By courtesy of New York Zoological Society

Sand shark (*Odontaspis taurus*).

The 27 species implicated by the Shark Research Panel in attacks on persons or boats are mostly large sharks with large, cutting teeth. Size, however, is not a dependable criterion, for man-eaters become dangerous when they are about one metre (three or four feet) long; and the largest ones, the basking shark and the whale shark, which grow to 12 and 18 metres (40 and 60 feet), respectively, subsist on minute planktonic organisms and on small schooling fishes. Although either might attack a boat if provoked, only two records of such occurrences have been reported, both in Scotland and both identified with the basking shark. More than 85 percent of all the shark species are too small, too unsuitably toothed, or too sluggish or live at depths too great to be potentially dangerous. The most dangerous sharks include in addition to the white shark, the hammerheads (*Sphyrna*), tiger (*Galeocerdo*), blue (*Prionace*), and sand sharks (*Odontaspis*).

Most stingrays live in shallow coastal waters. Some move with the tides to and from beaches, mud flats, or sand flats. Anyone wading in shallow water where these fishes occur runs some risk of stepping on one and provoking an instant response—the ray lashes back its tail, inflicting an agonizingly painful wound that occasionally leads to fatal complications. Rays can be serious pests to shellfisheries, for they are extremely destructive to oyster and clam beds.

NATURAL HISTORY

Food habits. All sharks are carnivorous and, with a few exceptions, have broad feeding preferences, governed largely by the size and availability of the prey. The recorded food of the tiger shark (*Galeocerdo cuvieri*), for example, includes a wide variety of fishes (including other sharks, skates, and stingrays), sea turtles, birds, sea lions, crustaceans, squid, and even carrion such as dead dogs and garbage thrown from ships. Sleeper sharks (*Somniosus*), which occur mainly in polar and subpolar regions, are known to feed on fishes, small whales, squid, crabs, seals, and carrion from whaling stations. Many bottom-dwelling sharks, such as the smooth dogfishes (*Triakis* and *Mustelus*), take crabs, lobsters, and other crustaceans, as well as small fishes.

The two giant sharks, the whale shark (*Rhincodon typus*) and basking shark (*Cetorhinus maximus*), resemble the baleen whales in feeding mode as well as in size. They feed exclusively or chiefly on minute passively drifting organisms (plankton). To remove these from the water and concentrate them, each of these species is equipped with a special straining apparatus analogous to baleen in whales. The basking shark has modified gill rakers, the whale shark elaborate spongy tissue supported by the gill arches. The whale shark also eats small, schooling fishes.

The saw sharks (Pristiophoridae) and sawfishes (Pristidae) share a specialized mode of feeding that depends on the use of the long, bladelike snout, or "saw." Equipped with sharp teeth on its sides, the saw is slashed from side to side, impaling, stunning, or cutting the prey fish. Saw sharks live in midwaters; sawfishes, like most other rays, are bottom inhabitants.

Thresher sharks (*Alopias*) feed on open-water schooling fishes, such as mackerel, herring, and bonito, and on squid. The long upper lobe of the tail, which may be half the total length of the shark, is used to frighten the fish (sometimes by flailing the water surface) into a concentrated mass convenient for slaughter.

Feeding methods of larger sharks

Most sharks and probably most rays segregate according to size, a habit that protects smaller individuals from predation by larger ones. Even among sharks of a size category, dominance between species is apparent in feeding competition, suggesting a definite nipping order. Other sharks keep clear of hammerheads (*Sphyrna*), whose manoeuvrability, enhanced by the rudder effect of the head, gives them an advantage. When potential prey is discovered, sharks circle it, appearing seemingly out of nowhere and frequently approaching from below. Feeding behaviour is stimulated by numbers and rapid swimming, when three or more sharks appear in the presence of food. Activity soon progresses from tight circling to rapid crisscross passes. Biting habits vary with feeding methods and dentition. Sharks with teeth adapted for shearing and sawing are aided in biting by body motions that include rotation of the whole body, twisting movements of the head, and rapid vibrations of the head. As the shark comes into position, the jaws are protruded, erecting and locking the teeth into position. The bite is extremely powerful; a mako shark (*Isurus*), when attacking a swordfish too large to be swallowed whole, may remove the prey's tail with one bite. Under strong feeding stimuli, the sharks' excitement may intensify into what is termed a feeding frenzy, in which not only the prey but also injured members of the feeding pack are devoured, regardless of size.

In most cases the initial attraction to the food is by smell. Laboratory studies have shown that sharks do not experience hunger in the normal sense of the word, and they are much more prone to be stimulated to feeding by the olfactory or visual cues announcing the appearance of prey.

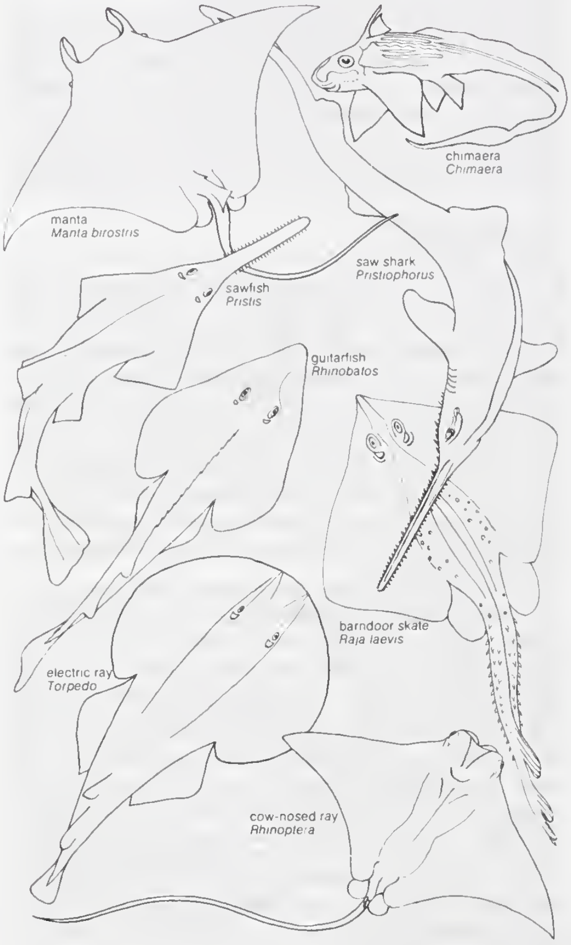


Figure 9: Body plans of representative selachii.

The majority of batoid fishes (members of the order Batoidei; *i.e.*, rays and allies) are bottom dwellers, preying on other animals on or near the sea floor. Guitarfishes (Rhynchobatidae and Rhinobatidae), butterfly rays (Gymnuridae), eagle rays (Mylobatidae), and cow-nosed rays (Rhinopteridae) feed on invertebrates, principally mollusks and crustaceans. Whip-tailed rays (Dasyatidae) use their broad pectoral fins to dig shellfish from sand or mud. Skates (Rajidae) lie on the bottom, often partially buried, and rise in pursuit of such active prey as herring, trapping the victims by swimming over and then settling upon them, a practice facilitated by the skates' habit of hunting at night.

Electric rays (Torpedinidae) are characteristically bottom fishes of sluggish habits. They feed on invertebrates and fish, which may be stunned by shocks produced from the formidable electric organs. With their electricity and widely extensible jaws, these rays are capable of taking very active fishes, such as flounder, eel, salmon, and dogfish. Shallow-water electric rays have been observed to trap fishes by suddenly raising the front of the body disk, while keeping the margins down, thereby forming a cavity into which the prey is drawn by the powerful inrush of water.

Most of the myliobatoid rays (seven recognized families of the suborder Myliobatoidea, which includes all of the typical rays) swim gracefully, with undulations of the broad, winglike pectoral fins. Some species, especially the eagle rays, frequently swim near the surface and even jump clear of the water, skimming a short distance through the air.

Manta, or devil, rays (Mobulidae) swim mostly at or near the surface, progressing by flapping motions of the pectoral fins. Even the largest often leap clear of the water. In feeding, a manta moves through masses of macroplankton or schools of small fish, turning slowly from side to side and using the prominent cephalic fins, which project forward on each side of the mouth, to fan the prey into the broad mouth.

Chimaeras and ghost sharks (Chimaeridae) dwell near the bottom in coastal and deep waters, to depths of at least 2,500 metres (about 8,000 feet). They are active at night, feeding almost exclusively on small invertebrates and fishes.

Reproductive behaviour. Mature individuals of some species of sharks segregate by sex, coming together only during the mating season, when the males, at least those of the larger, more aggressive species, stop feeding. Segregation is a behavioral adaptation to protect the females, one principal courting activity used by the male to induce cooperation of the female in mating being that of slashing her with teeth especially developed for that purpose. After mating, the sexes again separate. The pregnant females also tend to keep apart from the other females of like size. As the time of parturition approaches, the pregnant females move to particular areas, which presumably have properties of environment especially suitable as nursery grounds. When giving birth to their young, they stop feeding, and, soon after parturition is completed, they depart.

Care of the young

Nursery areas vary with species. Some sharks—*e.g.*, the bull and sandbar sharks—use shallow waters of bays and estuaries; the silky shark uses the bottom far out on oceanic banks such as the Serrana Bank in the western Caribbean. The Atlantic spiny dogfish (*Squalus acanthias*) bears its young mostly during the winter far out on the continental shelf of northeastern America almost two years after mating.

A few skates that have been observed mating may be characteristic of other rays. The male seizes the female by biting the pectoral fin and presses his ventral surface against hers while inserting one, or in some species, both claspers into her cloaca. Male skates have one to five rows of clawlike spines on the dorsal side of each pectoral fin. These are retractile in grooves of the skin and are used to hold the female during mating.

The eggs of skates in aquaria have been observed to be extruded in series, usually of two but sometimes one, with rests of one to five days between extrusions. A female of a European skate, *Raja brachyura,* laid 25 eggs over a 49-day period in the aquarium located at Plymouth, England.

Although the mating of chimaeroids has not been observed, it is generally presumed that the mode of copulation is similar to that of sharks and that the male's frontal spine and anterior appendage of the pelvic fins are probably used in securing the female. Two eggs are laid simultaneously, one from each oviduct. They are often carried for a relatively long period before being laid, several hours or even days, each protruding for the greater part of its length.

FORM AND FUNCTION

Distinguishing features. The elasmobranchs are fishlike vertebrates differing from bony fishes in many respects. The skeleton is composed of cartilage and, although partly calcified (especially in the vertebrae), lacks true bone.

Shark scales

There are five to seven fully developed gill clefts, opening separately to the exterior. Most sharks and all rays have an opening behind each eye, called a spiracle, which is a modified first gill cleft. The dorsal fin or fins and fin spines are rigid, not erectile. Scales, if present, are structurally minute teeth, called dermal denticles, each consisting of a hollow cone of dentine surrounding a pulp cavity and covered externally by a layer of hard enamel-like substances called vitrodentine. The scales covering the skin do not grow throughout life as they do in bony fishes, but have a limited size; new scales form between existing ones as the body grows. Certain other structures, such as the teeth edging the rostrum (beak) of sawfishes and saw sharks, the stinging spines of sting rays, and the teeth in the mouth, are structurally modified scales. The teeth, arranged in rows in the mouth, are not firmly attached to the jaws but are imbedded in a fibrous membrane lying over the jaws. When a tooth becomes broken, worn, or lost, it is replaced by one moving forward from the next row behind; at the base of the innermost row are rudimentary teeth and tooth buds that develop and move forward as needed. A spiral membranous fold (spiral valve) extends through the intestine of all sharks, rays, and chimaeras.

The rays differ externally from sharks in having the gill openings confined to the lower surface; the eyes of the rays are on the dorsal surface, and the edges of the pectoral fins are attached to the sides of the head in front of the gill openings. Some rays lack scales, and others are variously armed with thorns, tubercles, or prickles, all of which are modified scales; the tails of some have long, saw-toothed spines equipped with poison glands. In the sawfishes the snout is prolonged into a long, flat blade armed on either side with teeth. Some skates and a few rays have electric organs by which they can administer electric shocks to enemies or prey.

The chimaeras have only one external gill opening. In the adult the skin on each side of the head is smooth and lacks scales; the teeth consist of six pairs of grinding plates. The dorsal fin and spine are erectile. Like male sharks and rays, male chimaeras have claspers that serve to transfer sperm to the female, but, in addition, they have an erectile clasping device, the tantaculum, in front of each pelvic fin; most species have another such organ on top of the head.

Senses. Although sharks are often said to have a low order of intelligence, they, as well as rays and chimaeras, have survived successfully over a long period of geologic time. They are well equipped to locate prey and their own kind; to direct the course of their seasonal migrations; to discriminate specific localities; to respond to variations of temperature; to react to attractive or repelling substances in the water; and perhaps even to feel objects some distance away from them. They can see, hear, smell, taste, feel, and maintain their equilibrium. The roles of the sense organs have been studied in only a few species, principally sharks, and consequently remain imperfectly understood.

The sense of smell is highly developed and probably the principal means of locating prey and guiding the predator toward it. Given a favorable direction of current, sharks can detect incredibly minute concentrations—fractions of a part per million (*i.e.*, less than 1×10^{-6} parts)—of certain substances in the water, such as blood.

Although their eyes are structurally and functionally adapted for seeing, it is believed that their visual acuity in discerning the form and colour of an object is limited. The importance of sight relative to smell increases as a shark approaches its target.

The hearing apparatus, located in the auditory capsule of the cranium, includes a system of semicircular canals, which are responsible for maintaining equilibrium. Sharks seem to be remarkably sensitive to sounds of low frequency and to possess extraordinary faculty for directional hearing. Whether or not hearing is more sensitive than smell has not yet been established.

Sensory organs identified as taste buds are located on the floor, sides, and roof of the mouth and on the throat, as well as on the tongue. Experiments on several species of large sharks indicate that they do discriminate food types, preferring tunas, for example, to other fish species. Under some conditions, however, they become less fastidious, going into a feeding frenzy in which they attack anything, including others of their own kind.

Sensory organs located in the skin of all sharks, rays, and chimaeras receive a variety of information—vibrations of low frequencies, temperature, salinity, pressure, and minute electrical stimuli, such as are produced by another fish in the vicinity. These organs are located in the lateral line system (a series of sensory pores along the side), in groups of pores on the head (ampullar organs), and in pit organs distributed on the back, flanks, and about the jaws.

The urea retention habitus

Salt and water balance. Most marine vertbrates maintain lower concentrations of salts and other chemicals in their blood than are found in seawater, and so face a continuous problem of water loss to the environment, because of the tendency of water to move through membranes from regions of low salt concentration to regions of higher concentration. The marine cartilaginous fishes differ from almost all of the bony fishes (except the coelacanths and aestivating lungfishes) in being able to reabsorb in the renal (kidney) tubules most of their nitrogenous waste products (urea and trimethylamine oxide) and to accumulate these products in their tissues and blood, an ability termed the urea retention habitus. The concentration within the body thus exceeds that of the surrounding seawater, and water moves into the body with no expenditure of energy. When any of these fishes moves into freshwater, as many do, the urine flow to the outside increases; hence, the concentration of urea in the blood decreases. In the sawfish, for example, the increase of urine output is more than twentyfold; the blood urea concentration decreases to less than one-third the amount observed in marine forms. Purely freshwater elasmobranchs, such as the stingrays of the Orinoco and Amazon drainage systems, seem to lack the urea retention habitus.

Respiration. Sharks with spiracles take in some water through them, but they breathe chiefly by opening the mouth while expanding the mouth-throat (bucco-pharyngeal) cavity and contracting the gill pouches to close the gill slits. With the mouth closed, they contract the bucco-pharyngeal cavity while dilating the gill pouches, thus drawing the water over the gills where the exchange of oxygen and carbon dioxide takes place. Then, with the mouth still closed, they contract the bucco-pharyngeal cavity and gill pouches and open the gill slits to expel the water. Most of the rays, on the other hand, take in water chiefly through the spiracles; these then close by contraction at their anterior margins, which bear rudimentary gill filaments and a spiracular valve. Folds of membrane on the roof and floor of the mouth prevent the water from passing down the throat and direct it to the gill openings. Skates, which usually hold the lower surface of the head slightly above the bottom, inhale some water through the mouth; mantas, which have small spiracles and live near the surface, respire chiefly through the mouth. Skates, stingrays, guitarfishes, and angel sharks frequently reverse the direction of flow through the spiracles, apparently to clear them of foreign matter. Chimaeras take in water chiefly through the nostrils, keeping the mouth closed for the most part. The water reaches the mouth primarily through grooves leading there from the nostrils.

Reproduction and development. All species of sharks, rays, and chimaeras produce large, yolk-rich eggs. These are fertilized internally, for which the males are equipped with two copulatory organs called claspers along the inner edges of the pelvic fins. Each clasper has a groove for guidance of sperm. The few published descriptions of mating sharks and rays are probably characteristic of the entire group. The male grasps one of the female's pectoral fins with his teeth to hold her in position as he inserts a clasper through a cavity (cloaca) and into a tube (oviduct). Males of most species probably use only one clasper at a time. The sperm travel to the anterior end of the oviduct, where they fertilize the eggs. The eggs then move down the oviduct past the shell gland, where they are covered by a shell or capsule.

In oviparous (egg-laying) species, which include some of the sharks, probably all the skates, possibly some of the guitarfishes, and all of the chimaeras, the eggs are en-

Egg characteristics

veloped in a horny shell, usually equipped with tendrils for coiling around solid objects or with spikelike projections for anchoring in mud or sand. The egg cases of most species are more or less pillow-shaped; those of the horned sharks (Heterodontidae) are screw-shaped with a spiral flange. The eggs of chimaeras are elliptic, spindle-shaped, or tadpole-shaped and open to the exterior through pores and slits that permit entrance of water during incubation. An egg of the whale shark found in the Gulf of Mexico measured 30 centimetres (12 inches) long by about 14 centimetres ($5\frac{1}{2}$ inches) wide and was eight centimetres (three inches) thick. Protected by the shell and nourished by the abundant yolk, the embryo of an oviparous species develops for $4\frac{1}{2}$ to $14\frac{3}{4}$ months before hatching.

The majority of sharks and most, possibly all, rays other than the skates are ovoviviparous (*i.e.*, the egg hatches within the mother). In this case, the egg is first coated in the shell gland with a temporary membranous capsule that lasts only during early development. After emerging from its capsule, the embryo remains in the oviduct of the mother, nourished by the yolk sac to which it remains attached. Embryos of some ovoviviparous sharks, notably the porbeagle (*Lamna nasus*), mako (*Isurus oxyrinchus*), and sand shark (*Odontaspis taurus*), ingest yolks of other eggs and even other embryos within the oviduct of the mother after the contents of their own yolk sacs are exhausted. In the majority of ovoviviparous sharks and rays, organically rich uterine secretions provide supplemental nourishment, which is absorbed by the yolk sac and in many cases by appendages borne on its stalk. In some genera of rays, vascular filaments producing these secretions extend through the spiracles and into the digestive tract of the embryos.

Several shark species are viviparous—*i.e.*, the yolk sac develops folds and projections that interdigitate with corresponding folds of the uterine wall, thus forming a yolk placenta through which nutrient material is passed from the mother.

Growth. Growth of a few shark species has been measured or estimated by the differences in length at the times of tagging and recapturing specimens, by statistical analysis of length in systematically collected samples, by the space between concentric circles on the centra of the vertebrae, and by periodic measurements of specimens kept in aquariums. All studies indicate a slow growth rate. During the 10 years between birth and maturity, male Atlantic spiny dogfish grow an average of 47 centimetres (19 inches) and females 67 centimetres (26 inches). The Greenland shark (*Somniosus microcephalus*), which attains $6\frac{1}{2}$ metres (21 feet) or more (although rarely taken larger than about four metres [13 feet]), grows only about $7\frac{1}{2}$ millimetres (a little more than $\frac{1}{4}$ inch) per year. The annual growth increments of tagged juvenile whitetip reef and Galápagos sharks, both species that become at least $2\frac{1}{2}$ metres (eight feet) long, were found to be 31 to 54 millimetres (just over one to two inches) and 41 millimetres (about $1\frac{1}{2}$ inches), respectively. The Australian school shark (*Galeorhinus australis*) grows about 80 millimetres (three inches) in its first year and 30 millimetres (one inch) in its 12th year. By its 22nd year, it is estimated to be approaching a maximum length of 160 centimetres (just over five feet).

The disk of the eastern Pacific round stingray (*Urolophus halleri*) increases in width on the average from 75 millimetres (three inches) at birth to 150 millimetres (six inches) when mature, when 2.6 years old. In the next five years it grows about 60 millimetres (about $2\frac{3}{8}$ inches) more toward its maximum recorded width of 25 centimetres (10 inches) in males or 31 centimetres ($12\frac{1}{4}$ inches) in females. The males of European thornback rays (*Raja clavata*) are about 50 centimetres (20 inches) wide when they reach first maturity, about seven years after birth; females are 60 to 70 centimetres (24 to 28 inches) at first maturity, nine years after birth.

EVOLUTION AND CLASSIFICATION

Evolution. The earliest fossil remains of fishlike vertebrates are too fragmentary to permit tracing the modern fishes precisely to their origins. It is believed that the ancestral forms evolved during the Silurian Period (from about 430,000,000 to 395,000,000 years ago) in the upper reaches of streams. During the end of the Silurian and the beginning of the Devonian that followed, there appeared an exceedingly diverse group of armour-plated fishes with jawlike structures, paired fins, and bony skeletal tissue. Paleontologists refer to these extinct forms as a distinct class, Placodermi. Between the beginning and end of the Devonian (the latter about 350,000,000 years ago), the placoderms reached their peak in diversity and numbers and almost completely died out; only a few lingered another 10,000,000 years into the Early Mississippian (roughly, the Lower Carboniferous). During their flowering, the placoderms evidently gave rise to the Osteichthyes (the bony fishes) and the Chondrichthyes (the cartilaginous fishes). Even though the lines of evolution remain to be discovered, it seems quite clear that the two groups evolved independently, the Chondrichthyes appearing much later than the Osteichthyes.

Relationship to bony fishes

Although a few sharklike forms remained in fresh water, the vast majority soon invaded the sea, perhaps in response to the arid Devonian climate. There they adapted to life in salt water by evolving the urea retention habitus (see above *Salt and water balance*). Their cartilaginous skeleton, far from representing an evolutionary stage antecedent to the Osteichthyes, as was once believed, is more than likely degenerate rather than primitive. Possibly their precursors were the petalichthyids, a group of Devonian sharklike placoderm fishes that had ossified skeletons and well-developed fins.

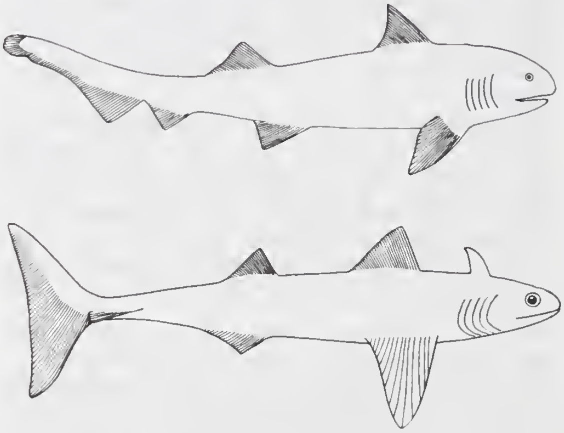


Figure 10: *Primitive sharklike fishes.*
(Above) *Hybodus* from the Mesozoic Era. (Below) *Cladoselache* from the Late Devonian Period.

The phyletic relationship of the chimaeras and the sharks and rays is a subject capable of varying interpretation. Although both groups have many characteristics in common, such as the possession of a cartilaginous skeleton, placoid scales, teeth simply embedded in gums, a spiral valve in the intestine, urea retention habitus, internal fertilization (for which the males have claspers), and the absence of a swim bladder, the two groups may have evolved independently along parallel lines, the chimaeras from the pyctodonts, an order of Devonian placoderms with body form and tooth structure very suggestive of modern chimaeras.

Fossil sharklike fishes

The first fishes clearly identified with the Chondrichthyes were sharklike in form. One order, the Pleurocanthodii, consisting of one family of freshwater, sharklike fishes, appeared in the Late Devonian, was abundant in the Carboniferous and Early Permian (until about 250,000,000 years ago), and disappeared during the Triassic Period, which followed. These fishes were characterized by the following features: the skeletal structure of both pectoral and pelvic fins had an axis with side branches (called the archipterygial type); the tail was almost symmetrical, being only slightly tilted upward; a long movable spine projected backward from the back of the head; the teeth had two divergent prongs and a central cusp set on a buttonlike base; the anal fin was two-lobed; and the males had claspers.

The other order, Cladoselachii, consisted of marine fishes known only from fossils of the late Middle Devonian, Carboniferous, and Early Permian periods. Their distinguishing characteristics were that each tooth had a long base composed of a bonelike tissue, from which rose three conical cusps, a tall central one and two smaller ones on either side; the body scales had several lobes or cusps; the jaws had double articulation, extending forward to the snout; claspers were lacking; the outline of the caudal (tail) fin was almost symmetrical but with differing internal structure of the upper and the lower lobes.

The cladoselachians were probably ancestral to a group closer to modern sharks, the order Hybodontii. They probably represent an intermediate state in selachian evolution and are classified by some authorities in the order Selachii. Although the jaws had the primitive double articulation, the skeletal support of the pectoral and pelvic fins was close to that of modern selachians, with basal elements projecting outward into the fins. The teeth near the front of the mouth were generally sharp-cusped; the cusps of those further back were sometimes reduced to a rounded crown. The front teeth were suitable for seizing prey; those in the back were suitable for crushing mollusks. The hybodonts appeared toward the end of the Devonian, flourished in the Late Paleozoic, and died out during the latter half of the Mesozoic, a few lasting into the Late Cretaceous (about 80,000,000 years ago).

The great period of radiation (diversification) in marine vertebrates characterizing the Mesozoic ended in the Permian, and the chondrichthyed fishes, which had reached their greatest flowering during the Carboniferous, became greatly reduced, remaining so until the Jurassic (about 190,000,000 years ago), when the areas of the seas expanded and those of the land diminished. Then the six-gilled shark (*Hexanchus*), horned shark (*Heterodontus*), and guitarfishes appeared. By the end of the Cretaceous, most of the families and many genera of modern sharks, as well as those of skates and rays, were represented. The evolution of elasmobranch fishes, much as they are known today, had been accomplished.

Annotated classification. The most recent approaches to a comprehensive review of the chondrichthyeds are that of the American ichthyologists H.B. Bigelow and W.C. Schroeder and that by the American paleontologist Alfred S. Romer. The following synopsis, based on their work, provides principal identifying characteristics of all major Recent groups.

CLASS CHONDRICHTHYES (or SELACHII)

Subclass Elasmobranchii (sharks and rays)

Chondrichthyeds with 5–7 pairs of gill clefts not covered by a fold of skin, opening separately to the exterior.

Order Selachii (sharks)

Elasmobranchs with gill clefts opening at least partly on the side of the body.

Suborder Notidanoidei. Sharks having 6 or 7 gill openings. Anal fin present.

Family Hexanchidae (cow shark and 7-gilled sharks). Lower Jurassic to present; marine. The cow shark (*Hexanchus griseus*), in deep water, down to 1,875 m (about 6,000 ft). Distinguished by presence of 6 gill slits; teeth of lower jaw strikingly unlike those of upper, the 5 or 6 on either side of the central tooth being about twice as broad as high, their inner edges saw-toothed with 5–8 pointed cusps. Size up to at least 5 m (about 16½ ft), estimated length at maturity about 2 m (about 6½ ft). Ovoviviparous; 4.5-m (15-ft) specimen contained 108 embryos. The 7-gilled sharks (*Heptranchias* and *Notorhynchus*) are widely distributed in warm and temperate continental waters.

Suborder Chlamydoselachoidei

Family Chlamydoselachidae (frilled shark). Miocene to present. One modern species known, rather rare. Distinguished by 6 gill slits, the margins of the first being continuous across the throat. Size to about 2 m (about 6½ ft). Moderately deep water of the eastern North Atlantic from Portugal to Norway and in the North Pacific off California and Japan.

Suborder Heterodontoidei. Upper Devonian to present. Five gill openings on each side of body; anal fin present; 2 dorsal fins, each preceded by a spine. Marine.

Family Heterodontidae (Horned sharks, bullhead, Port Jackson shark). With 1 Recent genus and about 10 species. Oviparous; egg case screw-shaped; a double spiral flange extending from apex to large end. Teeth in upper and lower jaws alike, those in front incisor-like, those on sides much larger and molar-like. Bottom dwellers out to about 180 m (about 590 ft) depth. Australia, New Zealand, East Africa, East Indies, China, Japan, eastern Pacific, north as well as south. Not known in Atlantic or Mediterranean. Size up to about 1.4 m (about 4½ ft).

Suborder Galeoidei (typical sharks). Five gill openings on each side of body; anal fin present; dorsal fin or fins not preceded by spines.

Family Odontaspididae (formerly *Carchariidae;* sand sharks). Upper Jurassic to present. Marine. Caudal peduncle (narrow "stalk" of the tail) without lateral keels; with a distinct pit on its upper surface but none on its lower. Teeth large, slender, smooth-edged, lower eyelid without a nictitating membrane (a transparent extra eyelid). Development is ovoviviparous; maximum size varies with species, from around 2.8 to 6 m (about 9 to 20 ft). One recent genus (*Odontaspis*) recognized, with some 6 species, found in warm temperate and tropical coastal waters of all oceans. Frequent shallow water near shore; sluggish except when feeding.

Family Scapanorhynchidae (goblin sharks). Lower Cretaceous to present. Marine. One genus, known from Japan, Portugal, and India, perhaps from Australia. Prominent elongation of the snout; protruding jaws. Maximum size to about 3.4 m (about 11 ft). Probably ovoviviparous. A deepwater shark, fished commercially in Japan for its liver and flesh.

Family Isuridae. Upper Cretaceous to present. Three genera, marine, although at least 1 species (the white shark) occasionally strays into estuaries. Distinguished by 2 dorsal fins, of which the first is much larger than the second and the rear end of its base situated well in advance of the pelvic fins; caudal fin lunate (crescent-shaped), its axis steeply raised. Teeth large. Ovoviviparous or viviparous. Circumglobal, occurring in boreal to warm temperate belts of all oceans in both hemispheres. Size in the great white shark (*Carcharodon carcharias*) varies from 1.4 to 6 m (4.6 to 19.7 ft) in length, but individuals may possibly exceed 8 m (26.2 ft). Three genera, *Lamna, Isurus,* and *Carcharodon.* the last 2 dangerous to man, the great white shark unquestionably the most dangerous of all fishes.

Family Cetorhinidae (basking shark). Oligocene to present. Marine. Two dorsal fins, the first well in advance of pelvics; lunate caudal fin; gill openings extending around sides almost meeting at throat. Hundreds of minute teeth. Ovoviviparous. Embryonic development undescribed. Size at birth probably 1.5–1.8 m (5–6 ft); maximum size to 13–14 m (42½–46 ft). Single genus (*Cetorhinus*) inhabiting temperate and boreal zones around the world. Whether basking sharks of the Northern and Southern hemispheres belong to a single species (*C. maximus*) is undetermined. Sluggish, inoffensive sharks, living at or near the surface, feeding wholly on plankton, which they sieve out of the water with their gill rakers.

Family Alopiidae (thresher sharks). Eocene to present. One genus, 5 species. Distinguished by the elongated upper lobe of the tail fin, which is almost as long as the rest of the body. Teeth small, bladelike. Ovoviviparous. Total length to about 6 m (20 ft). Cosmopolitan at low and middle latitudes of all oceans. Harmless to man. Occasionally sold for food.

Family Orectolobidae (carpet and nurse sharks, wobbegongs). Upper Jurassic to present. Marine. Distinguished by the presence of 2 dorsal fins, the origin of the first over or behind the pelvic fins; nostril connected with mouth by a deep groove, its anterior margin with a well-developed fleshy barbel (tentacle). Teeth small, with several cusps; development ovoviviparous in some, oviparous in others. Some species (carpet sharks) live on the bottom and are ornamented with fleshy flaps along the sides of the head. Large family of many genera and species occurring mostly in western Pacific, Australasia, Indian Ocean, Red Sea. Only 1 species, the nurse shark, in Atlantic.

Family Rhincodontidae (whale shark). Distinguished from all other sharks by large, lunate tail, mouth at end of snout, 3 prominent ridges extending the length of body along the sides, back marked with round white or yellow spots and a number of white or yellow transverse stripes. Oviparous. Size said to reach over 18 m (59 ft), the largest of modern fishlike lower vertebrates. One species only (*Rhincodon typus*); open waters of all oceans, mostly in tropics, but north to 42° N latitude (near New York) and south to 33°55′ S (Table Bay, South Africa). Sluggish and inoffensive.

Family Scyliorhinidae (cat sharks, European dog shark, swell sharks). Upper Jurassic to present. Most with 2 dorsal fins (1 genus with 1); first dorsal fin situated far back on body, at least half of it behind the origin of the pelvic fins. Furrows are more or less developed at the angle of the jaws; teeth small, numerous, with several cusps. A large group of small sharks comprising many genera, occurring in temperate to tropical latitudes. The swell sharks (*Cephaloscyllium*) can inflate the belly with air or water, presumably a defense mechanism. Of little, if any, commercial value; harmless to man.

Family Pseudotriakidae (false cat sharks). Distinguished by the base of the first dorsal fin being at least as long as the caudal fin. Teeth minute, numerous. One genus, *Pseudotriakis;* 2 species, 1 on both sides of the North Atlantic, the other in the western Pacific. Size to nearly 3 m (about 10 ft). Deepwater sharks (taken down to 1,477 metres [4,850 ft]) rarely straying near shore and known only from a few specimens.

Family Triakidae (smooth dogfishes). Upper Cretaceous to present. The principal distinguishing feature is small, closely crowded teeth in series, rounded or somewhat compressed and with 3 or 4 cusps. True nictitating membrane lacking in eye. Development ovoviviparous or viviparous. Small sharks of coastal waters in tropical to temperate zones of all oceans. The family comprises at least 7 genera and numerous species. Smallest species, *Triakis barbour,* reaches only about 40 cm (16 in.); maximum size for others of family 150–175 cm (59 to 69 in.). Although sharks of this family are generally considered harmless, there is one authenticated case of a California leopard shark (*Triakis semifasciata*) attacking a man in northern California.

Family Carcharhinidae. The largest family of sharks, with 13 genera and numerous species, including the tiger shark, the great blue, whalers, and many with various local common names. Upper Cretaceous to present. Two dorsal fins, the first in front of the pelvics. All species except 1 with well-developed nictitating membrane. Teeth bladelike, with only 1 cusp, only 1 or 2 rows functional along sides of jaws. Development either ovoviviparous or viviparous. The species range in maximum size from about 1.4–5.5 m (about 4½ to 18 ft). Members of this family occur from tropical to temperate zones in all oceans. Although most species are marine, several frequent brackish water or freshwater, and some occur in lakes that connect with the sea. The *Carcharinus leucas-gangeticus* group, a collection of several closely related species or subspecies, has a bad reputation; several cases of unprovoked attacks on persons are on record in both salt water and freshwater.

Family Sphyrinidae (hammerhead sharks). Upper Cretaceous to present. The most obvious distinguishing feature is the lateral expansion of the head in a hammer or bonnet form, with the eyes at the outer edges. Teeth large, triangular, smooth edged in some species, serrate in others. Viviparous or ovoviviparous; size varies with species, the largest (*Sphyrna mokarran*) is said to reach 6 m (about 20 ft). Predacious. Marine, but occasionally straying into estuaries. Occur in tropical and temperate zones of all seas. Hammerheads have a sinister reputation of initiating unprovoked attacks, documented by authoritative cases on record.

Suborder Squaloidei (spiny dogfishes, bramble sharks, sleeper sharks, pygmy sharks). Upper Cretaceous to present. Widely distributed, found in all of the oceans from tropical to both Arctic and sub-Antarctic latitudes; from shallow to deep depths. Anal fin lacking; snout not elongated into a beak; body subcylindrical (nearly round in section); not flattened dorsoventrally; margins of pectoral fin not expanded forward past first pair of gill openings.

Family Squalidae (spiny dogfishes, sleeper sharks and several others lacking common names). Upper Cretaceous to present. Distinguished by having about as many upper teeth in anterior row as in succeeding rows. Diverse forms, habits, and sizes. Spiny dogfishes (*Squalus*) grow to about 120 cm (47¼ in.); the Greenland sleeper shark to over 6 m (about 20 ft); a pygmy shark (*Euprotomicrus*) to about 26 cm (10¼ in.). Sleeper sharks (*Somniosus*) taken for food in waters around Iceland and west Greenland, but the fish must be dried before eating; otherwise it produces a mild poison.

Family Oxynotidae (prickly dogfish). Miocene to present. Distinguished by number of functional upper teeth increasing in each row from front to rear; dermal denticles large and prominent. Taken from depths of 60–530 m (about 200 to 1,740 ft); 2 species known in eastern North Atlantic, Tasmania, and New Zealand.

Suborder Pristiophoridei

Family Pristiophoridae (saw sharks). Cretaceous to present. Anal fin lacking, snout greatly elongated, each edge studded with sharp toothlike structures; upper eyelid is free; gill slits at the side of the head, not underneath as in the sawfish. Ovoviviparous. Marine. Indo-Pacific, South Africa, Tasmania, Australia, Philippines, Korea, Japan. The order comprises 1 family, 2 genera, *Pristiophorous,* with 5 gill openings, and *Pliotrema,* with 6. Good food fish, harmless to man.

Suborder Squatinoidei

Family Squatinidae (angel sharks). Upper Jurassic to present. Marine, widely distributed in continental temperate and warm waters of Atlantic and Pacific oceans, on or close to the sea bottom. Characterized by flattened body, eyes on upper surface; anterior margin of pectoral fins far overlapping gill openings, which are partly on side of body; no anal fin. Largest up to about 2.4 m (about 8 ft). Ovoviviparous. One genus; possibly as many as 11 species.

Order Batoidei (rays, sawfishes, guitarfishes, skates, and stingrays)

Jurassic to present. Five gill openings, wholly on ventral surface; pectoral fins united with sides of head forward past the gill opening. Differ from all sharks in lacking upper free eyelid.

Suborder Pristoidei

Family Pristidae (sawfishes). Jurassic to present. Distinguished by extension of snout into long, narrow, flattened blade armed on either side with teeth but without barbels; gills on lower side of body, as in other batoids. Ovoviviparous. Size varies with species; common Atlantic sawfish to at least 5.5 m (18 ft); species in Indian and Australian waters to over 7 m (23 ft). Widely distributed in tropical and subtropical zones of all oceans; occur in estuaries and run far up large rivers into freshwater; but whether they remain resident and reproduce in freshwater lakes is not clearly established. Six species are known.

Suborder Rhinobatoidei (guitarfishes). Lower Jurassic to present. Electric organs are lacking; well-developed dorsal and caudal fins are present; base of tail is stout, not sharply marked off from rest of body. Most species are ovoviviparous, some perhaps oviparous.

Family Rhynchobatidae. Cretaceous to Recent. Distinguished by caudal fin being conspicuously bilobed and somewhat lunate; posterior edge of pectorals does not reach fore-margin of pelvics. Two genera, widely distributed in tropical and subtropical shallow waters of Indo-Pacific. Maximum size over 2 m (6½ ft).

Family Rhinobatidae. Caudal fin not bilobed; posterior edges of pectoral fins extending rearward at least as far as the origin of the pelvics. Small, rounded, closely set teeth. About 7 genera and 26 species; tropical and warm temperate shallow coastal waters of all oceans, in some localities entering freshwater and perhaps even permanently residing and breeding there. Size to about 1.8 m (about 6 ft). Harmless to bathers.

Suborder Torpedinoidei (electric rays, numbfishes, torpedoes). Eocene to present. Distinguished principally by highly developed electric organs on either side of the head and gill chambers; the outlines of these organs visible externally in most species. Pectoral fins with the head form a circular or ovate disk. Skin of most species soft and entirely scaleless. Eyes small, fuctional in most species but rudimentary or obsolete in deepwater forms. Mostly sluggish bottom dwellers in all the oceans from tropical to temperate latitudes and from the intertidal zone to depths of at least 1,100 m (3,600 ft). Three families, Torpedinidae, Narkidae, and Temeridae, distinguished by whether 1, 2, or no dorsal fins are present. Numerous genera and species. The largest electric rays of the genus *Torpedo* reach a length of about 180 cm (71 in.); the smallest, of the genus *Narke,* less than 30 cm (about 12 in.).

Suborder Rajoidei (skates). Lower Cretaceous to present. Moderately slender tail, on which the caudal fin is reduced to a membranous fold, though sometimes the caudal fin is entirely lacking; outer margins of the pelvic fin are more or less concave or notched. It is probable that all of the species are oviparous. Three families are distinguished by whether 1, 2, or no dorsal fins are present.

Family Rajidae (the great majority of skates). Two dorsal fins. Upper surface of the body disk more or less rough with spines, thornlike denticles, or both. Some species with electric organs along the sides of the tail, which, as far as known, produce very weak shocks. Six genera, widely distributed from tropical to subarctic belts of both hemispheres but with curious gaps in distribution; scarce, if present, in the Micronesian, Polynesian, and Hawaiian islands in the Pacific, in the western Atlantic between Yucatán and mid-Brazil, and in West Africa between Cape Verde and Walfish Bay. They occur from estuaries seaward, several species down to depths of over 500 m (1,640 ft). Several species inhabit deep water, at last one being found at over 2,700 m (almost 9,000 ft). They live mostly on the bottom, often partially buried.

Family Arhynchobatidae. Distinguished from other skates by having a single dorsal fin. Single genus and species, *Arynchobatis asperimus,* known only from New Zealand.

Family Anacanthobatidae. No dorsal fin; completely smooth skin; the pelvic fins so deeply notched as to form leglike structures anteriorly. Two genera, *Anacanthobatis* from Natal coast, South Africa, and *Springeria* from Gulf of Mexico.

Suborder Myliobatoidei. Upper Cretaceous to present. Distinguished by a slender tail, usually whiplike toward the tip; outer margin of the pelvic fins being straight or convex. Most with 1 or more saw-toothed, poisonous spines on upper surface of tail. Seven families are recognized. Tropical to warm temperate waters of all oceans, most abundant in shallow depths, entering brackish water and freshwater freely. One family is confined to freshwater.

Family Dasyatidae (whip-tailed rays). Lower Cretaceous to present. Caudal fin lacking; no distinct dorsal fin; tail, measured from the anus to the tip, longer than the breadth of the disk. Ovoviviparous. Tropical to warm temperate latitudes in all oceans. Generally in depths less than about 100 m (328 ft), most abundant close to shore, including tidal embayments. The largest reaches at least 2 m (6½ ft) in breadth. Five genera, 2 in tropical and subtropical rivers of South America. A peculiarity in the structure of the pelvis has been used to differentiate a separate family, Potamotrygonidae.

Family Gymnuridae (butterfly rays). Miocene to present. Distinguished by the body being more than 1.5 times as broad as long and the tail considerably shorter than the body. Saw-toothed spine on the back of the tail in some species but not all. Maximum breadth about 2 m (6½ ft). Shallow coastal waters of tidal embayments and river mouths in tropical to warm-temperature latitudes of all oceans.

Family Urolophidae (stingrays). Eocene to present. Distinguished by having well-developed tail fin supported by cartilaginous rays; tail with at least one large saw-toothed spine. Ovoviviparous. The numerous species look very much alike; the largest does not exceed about 70 cm (27½ in.) in breadth. Tropical to warm temperate coastal waters less than about 70 m (230 ft) deep in western Atlantic and both sides of the Pacific from Japan to Tasmania, including the East Indies; they are unreported from eastern Atlantic or the Indian or African coasts of the Indian Ocean.

Family Myliobatidae (eagle rays). Upper Cretaceous to present. Distinguished from other myliobatoids by the forepart of the head projecting conspicuously beyond the rest of the body; eyes and spiracles on the sides of the head; tail as long as the disk or much longer and in most species bears a serrate venomous spine. Ovoviviparous. Some attain a width of about 2.5 m (about 8 ft). Cosmopolitan, occurring in continental waters and around islands and island groups from tropical to temperate latitudes; 4 genera.

Family Rhinopteridae (cow-nosed rays). Upper Cretaceous to present. Similar to eagle rays except that the projecting head is deeply incised at the midline, forming two distinct lobes. Ovoviviparous. Maximum breadth about 2 m (about 6½ ft). Coastal waters of tropical and warm temperate latitudes of all oceans.

Family Mobulidae (devil rays, or mantas). Pliocene to present. Continental waters and around offshore island groups of tropical to warm temperate belts of all oceans. Distinguished by a pair of armlike structures (cephalic fins) projecting forward, one on each side of the head. Tail whiplike; with or without a serrate edged spine. Teeth minute, arranged in many rows. Maximum size (breadth) of smallest species about 60 cm (about 24 in.); largest species at least 7 m (23 ft).

Subclass Holocephali (chimaeras, ghost sharks)

Upper Devonian to present. Cartilaginous skeleton, 4 pairs of gills, covered on each side of the body by an opercular fold of skin leading to a single external gill opening. First dorsal fin and spine erectile. Skin with small denticles along midline of back in some species and on tentacula and claspers of males. Teeth united to form grinding plates. Claspers of males are supplemented by an erectile organ, called tentaculum, in front of the pelvic fins, and all except one genus (*Harriotta*) have another club-shaped tentaculum on the forehead. Oviparous, laying elliptical, spindle-shaped, or tadpole-shaped eggs enclosed in brown horny capsules, remarkably large in proportion to the size of the parent. In breathing, chimaeroids take in water chiefly through the nostrils and thence through grooves leading to the mouth, which is generally kept closed. Variously distributed in temperate and boreal zones of all oceans, in coastal waters, and river estuaries and seaward down to over 2,500 m (8,200 ft).

Order Chimaerae

Family Chimaeridae (ghost sharks, ratfishes, chimaeras). Lower Jurassic to present. Rounded short or conical snout. Claspers of males bifid or trifid. Size to about 1.5 m (about 5 ft). Warm temperate and boreal latitudes of all oceans. Two genera, each with several species.

Family Callorhinchidae (elephant fish). Hoe-shaped proboscis. One genus (*Callorhinchus*) with a few species, which may eventually prove to be identical. Size to about 1.3 m (4 ft 3 in.). Restricted to cool temperate and boreal latitudes of Southern Hemisphere, generally taken in rather shallow water, sometimes entering estuaries and rivers.

Family Rhinochimaeridae (long-nosed chimaeras). Snout projecting into a long, straight point. Lateral line an open groove. Size to about 1.3 m (4 ft 3 in.). Probably cosmopolitan in middle latitudes of both hemispheres, taken in depths of 685–2,000 m (2,250 to 6,560 ft).

Critical appraisal. Many of the elasmobranchs are difficult subjects for taxonomic study. Differences between species are often subtle and hard to measure. Lacking skeletal support such as that possessed by the bony fishes, captured sharks collapse along the soft undersides of the body when taken out of the water, thus reducing the accuracy of measurements. A satisfactory taxonomic study of any species requires adequate samples over a full range of sizes, representing the full geographical distribution of the species. The sampling allow for rather large variations in body proportions between individuals of like size and different size groups and between populations inhabiting different regions of the total distribution. Hence, the identity of many species remains unsettled. The number of living species of sharks, now estimated at 200 to 250, tends to diminish as ichthyologists in different parts of the world accumulate and exchange careful anatomical measurements of fresh specimens, discovering that fishes from widely separated areas, formerly thought to be distinct, are actually of the same species.

The rays, except the larger ones, are somewhat easier to work with. About 300 to 340 species have been described. Here again, however, the number tends to diminish as comparative studies in different parts of the world show many of them to be cosmopolitan.

The classification of chondrichthyed fishes is a somewhat controversial subject. An authoritative opinion as to how sharks, rays, and chimaeras should be grouped can be reached only from a comprehensive critical review of all available pertinent living and fossil material. Students continuously add to the accumulation of field measurements and museum specimens, and so such a classification needs to be revised from time to time. Because this revision involves a vast amount of work, it is not undertaken often. (L.A.Wa.)

The fleshy-finned fishes (Sarcopterygii)

COELACANTHS (CROSSOPTERYGII)

The Crossopterygii constitutes a largely extinct subclass of primitive bony fishes that appeared at the beginning of the Devonian Period (395,000,000 years ago) but are now represented only by the coelacanth (*Latimeria chalumnae*).

General features. One major trait of the subclass is the division of the skull into an anterior, or ethmosphenoidal, unit and a posterior, or oto-occipital, unit, similar to the two cartilaginous prototypes found in the embryonic cranium. A strong joint unites the two regions at each side. The base of the skull and the vertebral column, which are incompletely ossified, allow the persistence, to various degrees, of the initial skeletal axis, or notochord. The subclass comprises three orders: Rhipidistia, Actinistia, and Struniiformes. Some authorities consider the Crossopterygii to be an order and what are treated here as orders to be suborders. After being widely distributed around the world in the Mesozoic Era, which began about 225,000,000 years ago, the crossopterygians underwent a rapid decline and then almost became extinct after the Triassic Period (190,000,000 years ago).

The Rhipidistia, predatory fishes of the Mesozoic, were ancestral to the terrestrial vertebrates, lived in freshwater, and probably had two respiratory apparatuses, a branchial (gill) system for aquatic respiration and a pulmonary (lung) system for air breathing. To facilitate air breathing the nasal cavities were provided with posterior nares (nostrils) homologous with the primary choanae (internal openings to the pharynx) of more advanced vertebrates. The skeletal structure of the paired fins clearly shows the ability for locomotion both on solid ground and in the water. The rhipidistians are thus credited, in the history of vertebrate evolution, with having made the great transition in anatomy and physiology involved in the emergence from water and resulting in the evolution of the amphibians. A Swedish paleontologist, E. Jarvik, has suggested that the rhipidistian Osteolepiformes gave rise to the Stegocephalia (the extinct ancestors of reptiles, birds, and mammals) and the Anura (frogs) and that the suborder Porolepiformes gave rise to the Urodela (salamanders).

Importance in the evolution of land animals

The Actinistia, especially the family Coelacanthidae, un-

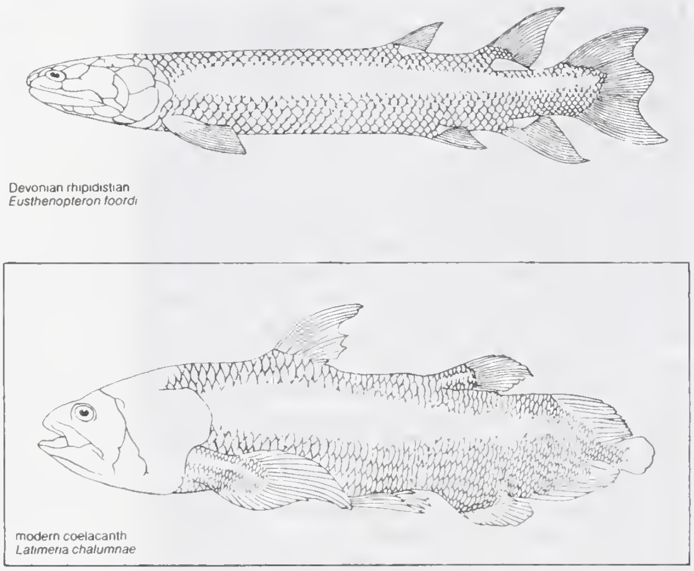


Figure 11: Representative crossopterygians.
Drawing by T. Kovacs based on (Top) *De tidiga fossila Ryggradsdjuren* (1959)

like the Rhipidistia, have exhibited exceptional evolutionary stability. The same fossil deposits contain both marine and freshwater types, both already specialized during the Devonian. They were thought to have disappeared 50,000,000 to 70,000,000 years ago, but in 1938 a live specimen was taken in the Indian Ocean. South African ichthyologist J.L.B. Smith identified it as a member of the Coelacanthidae and named it *Latimeria chalumnae,* the generic name in honour of Miss Courtenay Latimer, an associate who first brought the strange fish to his notice, the species name recalling its capture near the mouth of the Chalumna River. Between 1952 and 1970, more than 60 specimens of *Latimeria* were caught on the volcanic slopes of the Comoro Islands, at depths of 200 to 300 metres (650 to 1,000 feet).

Form and function. *Latimeria* has made it possible to reconstruct, with a high probability of accuracy, the anatomy of the coelacanths in general, in particular that of the perishable organs. Among the most striking characteristics are those of the head. The brain exhibits a relatively simple and harmonious structure, has an extremely small volume by comparison with the cranial capacity, and shows considerable displacement of the forebrain relative to the floor of the skull. The snout contains a special sensory organ, the rostral organ. At the intracranial articulation are attached some fibrous connective tissues as well as a pair of powerful longitudinal muscles, the subcranial muscles. This pair of muscles encloses the imposing notochord (a slender skeletal structure), the morphology of which must have been largely the same throughout the crossopterygian group. The heart of the *Latimeria,* which is very primitive, exhibits almost perfect bilateral symmetry (mirror-image form). It lies within a substantial pericardial cavity that retains the primitive continuity with the peritoneal (abdominal) cavity. There is a series of small valves near the exit from the heart, and several small contractile organs attached to the branchial arteries apparently fulfill the necessary function of assisting the propulsion of blood. On the whole, an embryonic condition co-exists with specialized arrangements.

An enormous cylinder of adipose (fat) tissue, aligned with a short median diverticulum (a blind pouch) of the ventral wall of the esophagus, lies above the abdominal organs. It apparently is a result of the degeneration of a lung apparatus. The extraordinary size of this cylinder is related to a displacement of the kidneys that undoubtedly occurs in the course of development, these organs occupying an unusual ventral position, posterior to the anus. Segments of the sympathetic nervous system are carried along in this movement.

Fins

The body is covered with large rough scales. The powerful tail fin has three lobes, lying in the median plane. The posterior end of the notochord extends into the middle lobe, which is by far the smallest. Two pairs of fins, the pectoral and the pelvic, are attached to their respective girdles. The base of each fin consists of a fleshy stalk supported by several successive segments of bone or cartilage that are not homologous with the similar parts of the Rhipidistia. Median fins similarly formed grow from the posterior part of the body, the posterior dorsal (above) and anal (below) fins. Finally, there is an anterior dorsal fin which, in contrast to the foregoing, is of ray-finned (actinopterygial) type—*i.e.,* lacking the fleshy, supportive stalk.

The Struniiformes, discovered only recently, lived in the Devonian. Their bony remains indicate considerable differences from both the Rhipidistia and the Actinistia. The fossil remains indicate, however, that they possessed the major characteristic of the subclass, the division of the cranium into an anterior and a posterior part.

(J.D.A./Ja.M.)

LUNGFISHES (DIPNOI)

The Dipnoi comprise an order of fishes that first appeared in the Lower Devonian Period (about 370,000,000–395,000,000 years ago) and that today is represented by only six species. Some authorities recognize only three extant species. Known as lungfishes, the extant species occur in Africa, South America, and Australia. They are especially interesting because of their characteristic body forms, their generally large size, their erratic distribution over the tropical regions of the earth, and their peculiar mode of life.

General features. *Economic importance.* The economic importance of the lungfishes is slight. Only in certain parts of Africa, because of their abundance and size, are they of any value to man as food. They are obtained from the mud of dried river bottoms. The South American lungfish, which is obtained in the same manner, is eaten locally.

Size range and distribution. Most species grow to substantial size. The Australian lungfish, *Neoceratodus forsteri,* attains weights of up to 10 kilograms (about 22 pounds) and a length of 1.25 metres (about 50 inches). Of the African lungfishes, the yellow marbled Ethiopian species, *Protopterus aethiopicus,* is the largest, growing to a length of two metres (about 80 inches). The South American species, *Lepidosiren paradoxa,* reaches a length of 1.25 metres (about 50 inches).

The distribution of the Dipnoi strikingly parallels that of the unrelated osteoglossomorph fishes, another freshwater group. The Australian lungfish occurs in a very small region of Australia: in the marshes of Queensland, along Burnett River and St. Mary's River. Four species (*Protopterus*) occur in Africa, where they are chiefly concentrated in the equatorial belt, but occur as far north as Senegal and as far south as Mozambique. Within their areas of distribution, the African protopterids are abundant along the riverbanks, in submerged areas with plant cover, and in lakes. *Lepidosiren paradoxa,* the South American lungfish, is widely distributed in that continent. It is especially numerous and often associated with the "eel" *Synbranchus marmoratus* in the shallow and muddy watercourses of the Chaco River in Paraguay and in neighbouring areas.

Drawing by J. Helmer from (*Lepidosiren, Protopterus*) J.R. Norman *A History of Fishes* (1947), Hill & Wang Publishers, (*Dipterus*) A.S. Romer *The Vertebrate Story,* Copyright 1933, 1939, 1941, 1959 by the University of Chicago

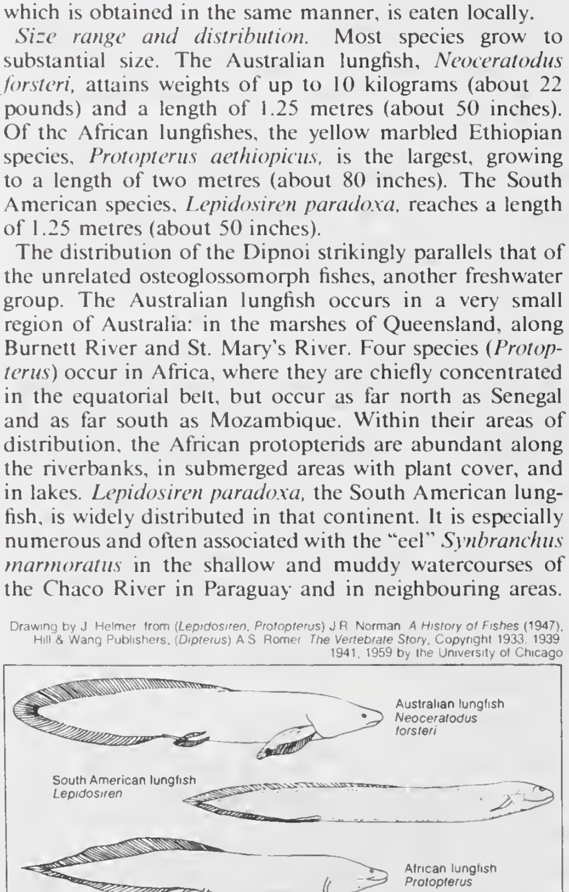


Figure 12: Living and fossil forms of Dipnoi fishes.

Natural history. *Reproduction and life cycle.* The African lungfishes spawn in the last half of winter, the onset of the rainy season. *Protopterus* species build a nest in the form of a pit on the bottom of a watercourse. The egg is about 3.5 to four millimetres (just over 1/8 inch) in diameter, and the tiny larvae emerge a week after the eggs are laid. The larvae have long, bright red, tuftlike or fanlike external gills, which they use for breathing until the lungs are fully developed. The young at first remain in the nest under the protection of the male.

The South American lungfishes dig a nest in the bottom in the form of a vertical passage, which frequently turns horizontally at the bottom. The male remains in the nest and guards the brood. During the spawning season, the pelvic fins of the male develop numerous tuft-shaped growths filled with small blood vessels (capillaries). These growths are believed to release oxygen from the blood, thereby oxygenating the water around the young.

The Australian lungfish lays gelatinous eggs among waterplants; the larvae, which have no external gills, breathe through internal gills.

Behaviour and ecology. Lungfishes are voracious, eating a variety of aquatic animals, including members of their own species. In captivity, African lungfishes eat earthworms, pieces of meat, tadpoles, small frogs, and small fish. The Ethiopian lungfish has at the front of the upper jaw two rather rounded teeth with a hard transverse (from side to side) bridge. The lower jaw has a number of crushing teeth. The prey is sucked in, crushed, and thoroughly chewed; such a manner of eating is rare among fishes.

Fins as sense organs

Form and function. *General features.* The slim, eel-like African protopterids and the even slimmer South American *Lepidosiren paradoxa* have long, stringy, very mobile pectoral and pelvic fins that are in a constant state of agitation—touching and sensing surroundings. The tips of these fins have a highly developed sense of touch, which, together with the fish's well-developed sensitivity to pressure and turbulence and its good sense of smell and taste, largely make up for the weakness of the eyes. The fish are almost blind with respect to the perception of form and movement. Pressure and turbulence are sensed by means of sensory structures called lateral lines. At the anterior, or head, end, the lateral lines are modified into a pattern of intricately interlaced bright lines, which are a series of tiny bud-shaped terminal organs. The highly individual patterns are used in distinguishing species.

The Australian lungfish has an entirely different appearance. It is more compactly built and has large, overlapping scales. The pectoral and pelvic fins are much broader. The African and South American lungfishes have paired lung sacs; in the Australian species the left lung sac atrophies.

Adaptations for breathing. There are a number of fishes that, in addition to or in place of gill breathing, have developed special organs through which they can breathe atmospheric air at the water surface. This occurs almost exclusively in freshwater fishes. In lungfishes these organs are, both in function and in structure, primitive lungs like those of amphibians. The name lungfish is thus well applied: these fishes have sac-shaped, pneumatic organs that lie along the alimentary tract. The inner surfaces of these air-breathing organs are covered with a great number of honeycomb-like cavities covered with fine blood vessels. As in terrestrial higher vertebrates, gas exchange takes place in the tiny air vesicles.

In order to breathe, the fish swims upward and positions its head so that the tip of the snout barely touches the water surface. The mouth is then opened wide, and the fish sucks in air from just above the water—a process often accompanied by a characteristic sound. The Australian lungfish reportedly breathes air through the nasal openings, the mouth remaining closed. In contrast to the higher bony fishes, lungfishes have a particular opening (choana) that connects the nasal cavity with the mouth.

In the Australian lungfish, gill breathing predominates at least some of the time—namely, in times of normal water level when the water is well oxygenated. At such times the fish rises less often to the surface to breathe atmospheric air. When the water level goes down, which usually occurs in August or September, the fish is often found in isolated waterholes in which the oxygen content is greatly reduced. Other fishes in such pools often die from lack of oxygen, but the lungfish survives, having changed over to the breathing of atmospheric air. During such a dry period the Australian lungfish surfaces about every 40 to 50 minutes for air. African lungfishes surface for air about every 30 minutes or, in some cases, at longer intervals.

The "dry sleep"

Physiology and biochemistry. African lungfishes bore into the bottom of a riverbed or lake bed for their "dry sleep." After burying themselves they become encased in a sheath that gradually hardens. Here they spend the dry season, during which the waterline becomes lower and the riverbed or lake bed finally dries out. The African lungfish generally digs in and encysts in this manner, even if there is sufficient time to swim to deeper waters. African lungfishes also burrow into mud and ensheath themselves under experimental conditions. They have been kept alive in such an induced state for more than two years.

The South American lungfish also bores into the mud in times of water shortage, but it forms no protective sheath. The Australian lungfish never buries itself in this manner.

Studies have shown that the "dry sleep" of the African lungfish is induced by a substance that inhibits the fish's normal metabolism.

Extracts from the brains of such sleeping fish injected into rats have caused them to become lethargic; in addition, the body temperature of the rats falls 5° C, and the metabolic rate falls 33 percent. The day after receiving such injections, the rats stop eating. It is believed that the substance responsible for this effect is a protein-like substance.

Evolution. The oldest Dipnoi, from the Lower Devonian Period, had skull and dental features that are characteristically dipnoid but also had many similarities to the crossopterygians (*e.g.*, coelacanth). The Dipnoi was abundant until Triassic times (190,000,000–225,000,000 years ago), after which their numbers decreased.

Dipterus, one of the oldest lungfish, had leaflike pectoral and pelvic fins similar to those of the modern Australian lungfish, and it seems reasonable to assume that early forms also had functional lungs comparable with those of species living today. Hardened sections of clay, cylindrical in shape, have been found in Pennsylvanian (about 280,000,000–325,000,000 years old) and Permian (225,000,000–280,000,000 years old) deposits. Remains of the dipnoid *Gnathorhiza,* closely allied to the extant African and South American species, were imbedded in the clay, strongly suggesting that they passed unfavourable conditions buried in mud.

An evolutionary line can be traced from *Dipterus* to *Neoceratodus,* the extant Australian genus. *Scaumenacia* and *Phaneropleuron,* common forms of the Upper Devonian (345,000,000–370,000,000 years ago), exhibited a much-reduced first dorsal fin (the first fin forward on the back); the second dorsal fin was enlarged and had shifted further toward the tail. Lungfish of Permian times showed an apparent fusion of the fins along the back and the rest of the vertical midline into the so-called diphycercal tail (*i.e.*, tapering to a point) present in modern lungfishes. Various side branches also occurred in the evolution of the Dipnoi, none of which has survived to modern times.

Classification. The annotated classification given below primarily relates to living forms; for a classification including the extinct forms see the critical appraisal below.

Distinguishing taxonomic features. The separation of Dipnoi as a discrete group is based largely on the structure and arrangement of the skull bones and the teeth. The suborders, of which there are two, are mutually distinguishable mainly by the number of lungs (one or two).

Annotated classification.

SUBCLASS DIPNOI

Lower Devonian (370,000,000–395,000,000 years ago) to Recent. Cranium not divided into movable parts; teeth in upper jaw reduced and lost in later members; some teeth fused into plates for eating shellfish. A single order.

Order Sirenoidei

Suborder Monopneuma

One functional lung.

Family Ceratidae. Pectoral and pelvic fins reduced but not tentacle-like; scales large; grows to about 1.25 m (about 50 in.); 1 living species: *Neoceratodus forsteri* (Australian lungfish).

Suborder Dipneuma
Two functional lungs.

Family Lepidosirenidae. Body eel-like in form; scales small, pectoral and pelvic fins modified into slender tentacle-like structures; passes dry periods in mud of dried river and lake bottoms; grows to about 2 m (about 80 in.); 2 living genera: *Lepidosiren* of South America (1 species: *L. paradoxa*) and *Protopterus* of Africa (4 species).

Critical appraisal. Some writers assign Dipnoi to the ordinal level, subsuming several families, mostly extinct, within that order.

The following alternate classification is according to A.S. Romer (1966), an American vertebrate paleontologist (extinct families represented only by fossils are indicated by a dagger [†]):

ORDER DIPNOI

†Family Dipnorhynchidae
Lower to Middle Devonian; Europe, Australia, North America.

†Family Dipteridae
Devonian; Europe, Greenland, North America, northern Asia, Australia.

†Family Ctenodontidae
Carboniferous (280,000,000–345,000,000 years ago); Europe, North America, Australia.

†Family Phaneropleuridae
Upper Devonian; North America, Greenland, Europe.

†Family Sagenodontidae
Mississippian to Lower Permian (250,000,000–280,000,000 years ago); North America, Europe.

†Family Uronemidae
Mississippian; Europe, North America (?).

†Family Conchopomidae
Pennsylvanian (280,000,000–325,000,000 years ago) to Lower Permian (250,000,000–280,000,000 years ago); North America, Europe.

Family Ceratodontidae
Lower Triassic (210,000,000–225,000,000 years ago) to Recent; one surviving species, *Neoceratodus forsteri.*

Family Lepidosirenidae
Pennsylvanian to Recent; 2 living genera, *Lepidosiren* and *Protopterus.* (K.H.L.)

THE HIGHER FISHES (ACTINOPTERYGII)

The early ray-finned fishes

STURGEONS, PADDLEFISHES, BICHIRS (CHONDROSTEI)

The Chondrostei comprise one of the three major subdivisions (subclasses, or infraclasses) of the class Actinopterygii, the higher, bony, ray-finned fishes. Fossilized chondrosteans first appear in rocks of the Middle Devonian Period (about 375,000,000 years ago), and some members of this subclass have persisted to the present time. The only living representatives are the sturgeons and paddlefishes; the living bichirs (polypterids and the closely related reedfish) of Africa are also considered to be chondrosteans by some ichthyologists.

Decline of the chondrosteans

The chondrosteans were most numerous and diversified during the last part of the Paleozoic Era (ending 225,000,000 years ago) and the beginning of the Mesozoic Era (beginning 225,000,000 years ago). With the rise of the holosteans and teleosts (the other two major subdivisions of the Actinopterygii) during the Mesozoic, the chondrosteans declined, until by the end of the Cretaceous Period (65,000,000 years ago) they had been reduced to a few genera.

The few living chondrosteans are highly specialized and aberrant forms. Their evolutionary history has not been clearly documented. Except for the sturgeon, which is a food fish for man and the source of caviar, they have no economic importance. A study of the living sturgeons, paddlefishes, and bichirs, however, provides some understanding of extinct forms.

General features. The chondrosteans are difficult to characterize as a group, but certain features are common to most of them. Generally the bony adult neurocranium, or braincase, is composed of two divisions, a larger ethmo-otic and a smaller occipital section. In the paddlefishes and sturgeons the braincase is mostly cartilaginous, with a few isolated areas of bone.

Drawing by J Helmer based on D.S Jordan, *A Guide to the Study of Fishes*

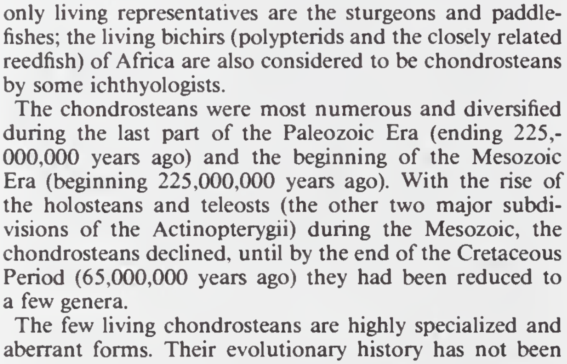


Figure 13: Body plans of modern chondrostean fishes.

Distribution and size range

The most numerous and widespread Paleozoic chondrostean fishes belong to the order Palaeonisciformes. The earliest known chondrosteans (Cheirolepidae and Stegotrachelidae), from the Middle Devonian of Europe, belong to this group. The palaeonisciforms inhabited a variety of freshwater and marine habitats and are known from all the continents with the exception of Antarctica. They reached their period of greatest number and diversity during the Carboniferous Period (280,000,000 to 345,000,000 years ago). Although the palaeonisciforms persisted with little modification into the Cretaceous, they began to show a marked decrease in numbers in the Triassic Period (190,000,000 to 225,000,000 years ago) and, by the end of the Cretaceous, had completely died out.

Modern sturgeons occur only in the waters of the Northern Hemisphere. The common sturgeon (*Acipenser sturio*) is found on the European coast from Norway to the Mediterranean Sea. A closely related form, probably of the same species, is found along the east coast of North America from the St. Lawrence River to the Gulf of Mexico. *A. gueldenstaedti* occurs in western Russia east to Lake Baikal. A smaller species, the sterlet (*A. ruthenus*), inhabits the Black and Caspian seas. *A. stellatus* occurs in rivers leading to the Black Sea, the Sea of Azov, and the Caspian Sea. The lake sturgeon of North America (*A. fulvescens*) occurs in the Mississippi valley, the Great Lakes, and northward into Canada. The white, Oregon, or Sacramento sturgeon (*A. transmontanus*) inhabits the waters of the Pacific Coast of North America from California to Alaska.

Bichirs (*Polypterus*) and the closely related reedfish (*Calamoichthys calabaricus*) occur in fresh waters of Central Africa. The Mississippi paddlefish (*Polyodon spathula*), also known as the spoonbill sturgeon, is found in the Mississippi basin; the Chinese paddlefish (*Psephurus gladius*), also called the swordbill sturgeon, occurs in the Yangtze River of China.

The length of some palaeonisciforms may have been as great as one metre (slightly more than three feet). The so-called subholosteans—a collective term for a heterogeneous group of chondrostean orders, from Perleidiformes through Parasemionotiformes (see below *Annotated classification*) probably grew to no more than 30 centimetres (about one foot) or so. Most modern sturgeons reach a length of little more than two metres (seven feet), but the hausen, or beluga (*Huso huso*), has been reported to reach 8.5 metres (28 feet). The Mississippi paddlefish grows to about 1.8 metres (about six feet), but the Chinese paddlefish sometimes reaches 6.3 metres (about 21 feet) in length. The largest species of bichir grows to about 70

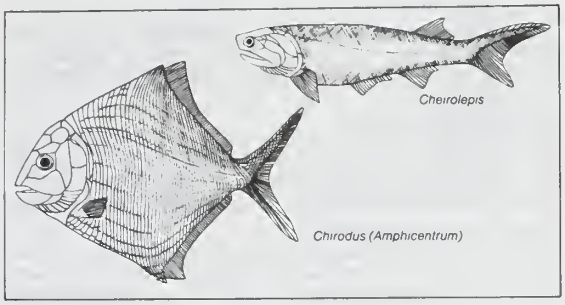


Figure 14: Examples of two chondrosteans of the order Palaeonisciformes.

From A.S. Romer, *Vertebrate Paleontology*, © 1933, 1945, 1966 by the University of Chicago, all rights reserved

centimetres (28 inches); the reedfish reaches a length of 90 centimetres (35 inches).

Natural history. *Reproduction and life cycle.* Marine sturgeons ascend rivers in spring or summer to deposit their spawn. They are abundant in the rivers leading to the Black and Caspian seas and to the Sea of Azov during the two weeks of the upstream migration. Early in summer the fish migrate into the rivers or toward the shores of freshwater lakes in large shoals for breeding purposes. The eggs are small and numerous, and the growth of the young is rapid. After the sturgeon attains maturity, growth continues at a slow rate for several years. Some attain great age: observations made in Russia indicate that the hausen may attain an age of 200 to 300 years.

Bichirs initiate courtship by leaping from the water. Little is known of their spawning habits. Young fish have external branching gills and are newtlike in appearance. Paddlefishes breed when seven or eight years old and spawn during spring floods. The larvae hatch in about two weeks and feed on their large yolk sac. The paddle, a long, broad extension of the snout, is absent at birth but begins to appear after two or three weeks.

Feeding

Ecology. Sturgeons occur in both salt water and fresh water. Ground feeders, they spend much time foraging, dragging their tactile, whiskerlike barbels over the bottom in search of small invertebrates and fishes. Paddlefishes feed by straining plankton (mostly tiny, drifting aquatic organisms) through their gill system and have been described as living plankton nets. Bichirs and reedfish mainly inhabit the edges of streams and floodplains. They remain concealed by day and forage at night for worms, insect larvae, crustaceans, and small fishes.

Form and function. *Extinct forms.* Most palaeonisciforms had fusiform (*i.e.,* tapered at both ends) bodies with blunt snouts, eyes situated far forward, pelvic fins located at about the middle of the body, dorsal (*i.e.,* back) and anal (on the lower side) fins nearly opposite one another on the posterior part of the body, and heterocercal (*i.e.,* with the top lobe longer than the lower lobe) caudal fins. With few exceptions, their bodies were covered with rhomboidal (diamond-shaped) scales, with or without a dentine layer. The scales articulated with one another by a peg-and-socket joint; in some groups, the scales tended to become thin and cycloidal, or rounded, as in the coccolepids. The rays of the unpaired fins were usually more numerous than their basal supports, and all the fins were usually bordered by scales that were generally larger and stronger than other scales (fulcral scales). A few families, such as the Late Paleozoic platysomids and amphicentrids, evolved deep, compressed bodies with elongated anal and dorsal fins.

In all palaeonisciforms, the upper jaw was tied to the cheekbones, which completely covered the area between the eyes and the gill covers. The jaw suspension may have had an oblique orientation (associated with a wide mouth gape) or a nearly vertical orientation (associated with a relatively smaller gape). The teeth either were rather well-developed or were sometimes practically absent. If present, they were generally styliform, or needlelike, in both the upper and lower jaws, and the musculature closing the mouth was rather straplike.

There is reason to believe that the biting mechanism in palaeonisciforms was less powerful than that of the holosteans. The arrangement of the fins and the structure of the tail suggest that manoeuvrability in swimming was not as great as in either the holosteans or the teleosts. Members of the Late Paleozoic order Tarrasiiformes had an elongated body and a diphycercal caudal fin that was continuous with the dorsal and anal fins. Haplolepiforms, also of the Late Paleozoic, had robust paired and unpaired fins and a relatively small number of unbranched fin rays. Like the palaeonisciforms, the subholosteans ranged from fusiform to deep-bodied.

In some subholosteans, the upper jaw was freed from the cheek elements and articulated with the skull only in the snout region. The palate and the cheek were also modified in such a way that the adductor, or closing, musculature of the lower jaw could enlarge to provide greater force in seizing prey. In connection with this, the upper border of the mandible developed an elevation (coronoid process) on its posterior part, and the attachment of part of the jaw musculature to this elevation increased the efficiency of the feeding mechanism. In the dorsal and anal fins, the number of fin rays tended to equal the number of basal supports, and the caudal fin became hemiheterocercal (*i.e.,* apparently symmetrical, but with the vertebral column turned upward and extending into the upper lobe).

Specializations for bottom feeding

Extant forms. The amount of bone in the sturgeon skeleton is less than that in the ancient forms. The modern sturgeon has bony plates on the head and five rows of bony shields along the body: one along the back, one on each side above the pectoral fins, and one on each side near the belly. The tail fin is heterocercal. The mouth is subterminal (*i.e.,* behind and below the snout tip), and this and other specializations are clearly related to bottom feeding. The mouth is toothless and is preceded by four fleshy barbels; the protractile lips have taste buds surrounding them. The form of the snout becomes more blunt and abbreviated with age.

The relationship of the paddlefishes to the sturgeons is not fully understood. The skeleton of the paddlefish, like that of the sturgeon, has lost much of its ossification. The body is fusiform, the fins well-developed, and the tail heterocercal. The elongated, paddle-shaped snout, which is composed entirely of cartilage, is one-third to one-half the total body length. The skin is smooth, except for a few scattered vestigial scales. The mouth is subterminal, and the jaw structure, particularly that of the adductor muscles, is suggestive of the palaeonisciform condition.

The bichir is rather elongated in form, the reedfish eellike; both have hard, diamond-shaped scales. The dorsal fin consists of a few to several separate finlets. The upper body is brown, grayish, or greenish, the lower side often white or yellowish.

Evolution. The long history of the chondrosteans, which extends over a period of 375,000,000 years, is marked by several important evolutionary events. The first is related to the appearance of the earliest ray-finned fishes, the palaeonisciforms. These fishes possess essentially the same feeding mechanism design and the same pattern, including a fully heterocercal tail, as later forms. The Late Paleozoic Tarrasiiformes and Haplolepiformes are obviously descended from the palaeonisciforms, but they are divergent enough to be regarded as separate orders.

The main groups of holosteans and halecostomes (which gave rise to the teleosts) apparently arose from subholostean-like ancestors during the Permian and Triassic periods. The heterogeneous subholosteans show modifications in the feeding mechanism and in the body that foreshadow the holosteans and halecostomes. Fishes referred to this unnatural group were characteristic of the Triassic Period, although a few families continued into the Jurassic (136,000,000 to 190,000,000 years ago). In general, the subholosteans can be said to show a diversity in the structure of the skeleton that was never attained by the more primitive palaeonisciforms. This diversity suggests the kind of evolutionary "experiments" that must have occurred during the rise of the various families of more advanced actinopterygians.

The origin of the order Acipenseriformes (which includes the sturgeon) is not known, although they were clearly derived from some palaeonisciform groups. Fossils that are without doubt related to the sturgeons and paddlefishes are no older than the Upper Cretaceous (about 65,000,000 to 100,000,000 years ago). Earlier, the history of this order is poorly documented and confused. Both the sturgeons and the paddlefishes became specialized early in their history and have shown only minor diversification since then.

The Polypteriformes, which include the living bichirs (*Polypterus*) and reedfish (*Calamoichthys*) of Africa, show a confusing array of palaeonisciform, holostean, and specialized characters. Some skull and scale features indicate derivation from palaeonisciform ancestors. The palate and jaws, on the other hand, suggest attainment of a nearly holostean-like pattern. However, the specialized fins, including the diphycercal tail, indicate that the polypteriforms have had a long, independent history. Fossil occurrences, which may extend back to the Early Tertiary Period (beginning 65,000,000 years ago), offer no clues to their affinity.

Classification. *Distinguishing taxonomic features.* The approximately 37 families of the Chondrostei are separated from one another, for the most part, on the basis of differences in dermal bone pattern, body shape, and fin form and position.

Two orders, Tarrasiiformes and Haplolepiformes, are quite palaeonisciform-like in many ways but diverge in other ways that clearly set them apart as separate categories.

An advanced group of some 12 orders of unrelated chondrosteans is popularly referred to as subholosteans. The different subholostean families possess various combinations of palaeonisciform and holostean characters, and they show a diversity in the structure of the skeleton that was never attained in the earlier palaeonisciforms.

Annotated classification. Groups marked with a dagger [†] are extinct and known only from fossils.

SUBCLASS (or Infraclass) CHONDROSTEI
A group that has undergone various evolutionary diversifications. The orders of the Chondrostei are specialized for certain habitats and ways of life, and many show trends toward the holostean–halecostome level of organization, especially in median fin structure and development of a hemiheterocercal tail.

†Order Palaeonisciformes
Lower Devonian to Middle Cretaceous. Mostly fusiform fishes with heterocercal tail; maxillary bone fixed; more fin rays than basal elements in the median fins; 37 families of wide distribution, early members freshwater, later marine.

†Order Tarrasiiformes
Carboniferous (about 280,000,000 to 345,000,000 years ago). Palaeoniscid-like, but with elongated body, a diphycercal tail, and dorsal and anal fins continuous with it. One family, Tarrasiidae; Scotland and Illinois.

†Order Haplolepiformes
Upper Carboniferous. Pecuiar fishes with stout, unbranched fin rays; large gular (*i.e.*, in the throat region) plates; small opercular (*i.e.*, gill cover) apparatus. One family, Europe and North America.

†Order Perleidiformes
Lower to Upper Triassic. With ganoid (*i.e.*, bony, diamond-shaped, and not overlapping) scales; fin rays equal number of basal supports rather than exceed them; tail hemiheterocercal. Three families; worldwide.

†Order Redfieldiiformes
Lower and Middle Triassic. Like Perleidiformes, but fin rays more numerous than basal elements in dorsal and anal fins. One family, Redfieldiidae, in freshwaters of South Africa, Australia, and North America.

†Order Dorypteriformes
Upper Permian (225,000,000 to 250,000,000 years ago). Deep-bodied, with very modified skull; scales confined to anterior part of trunk. One family, Dorypteridae; Europe, China.

†Order Bobasatraniiformes
Lower Triassic. Body deep, laterally compressed; fin rays slightly more numerous than basal supports; crushing dentition; pelvic fins absent. One family, Bobasatraniidae; marine; widely distributed.

†Order Pholidopleuriformes
Lower to Upper Triassic. Some relatively long and slender; dorsal and anal fins far back on body, origin of anal fin anterior to dorsal fin; fin rays more numerous than basal elements; tail hemiheterocercal; jaw support almost vertical or moderately oblique. One family, Pholidopleuridae; marine and freshwater; wide distribution.

†Order Peltopleuriformes
Upper Triassic. Large eyes; hemiheterocercal tail almost symmetrical externally; dentition weak. Two families, Peltopleuridae and Habroichthyidae; marine, perhaps some plankton feeding; Italy, China.

†Order Platysiagiformes
Lower Triassic to Lower Jurassic. Elongated, fusiform body; tail hemiheterocercal; rays of median fins probably equalled basal elements in number; teeth large, conical. One family, Platysiagidae; marine; probably predacious; Italy and England.

†Order Cephaloxeniformes
Middle to Upper Triassic. Body deep, fusiform; thick head bones and crushing dentition; tail hemiheterocercal. One family, Cephaloxenidae; marine, probably bottom-dwelling mollusc eaters; Italy.

†Order Luganoiiformes
Middle and Upper Triassic. Body fusiform; head somewhat flattened in the horizontal plane; some head bones fused; jaw suspension inclined forward; fin rays apparently equal to basal elements in number; tail hemiheterocercal. One family, Luganoiidae; marine; probably predacious midwater fishes; Italy.

†Order Ptycholepiformes
Middle Triassic to Upper Jurassic. Fusiform body; fin rays of median fins nearly equalling basal elements in number; jaw support almost vertical; teeth small. One family, Ptycholepididae; marine; presumably plankton feeders; Europe.

†Order Saurichthyiformes
Lower Triassic to Upper Jurassic. Elongate, slender; snout elongated; single dorsal fin far back on body, opposite anal fin; tail with nearly equal lobes; number of scale rows reduced, 1 dorsal, 1 ventral, and 1 along each side; jaw suspension almost vertical; teeth large, conical; jaws long. One family, Saurichthyidae; marine and freshwater; predacious; worldwide. Length about 7–150 cm (2¾ to 59 in.).

†Order Chondrosteiformes
Lower Triassic to Upper Jurassic. Body scales and skull bones reduced; snout moderately developed, maxillary and opercular bones reduced; jaw support somewhat inclined backward; median fins paleonisciform-like, rays more numerous than basal supports. Probably gave rise to sturgeons. One family, Chondrosteidae; marine; some were suctorial feeders like sturgeons; England.

†Order Parasemionotiformes
Lower Triassic. Near holosteans in dermal skull structure. Two families; marine; Siberia, Greenland, and Madagascar.

Order Acipenseriformes (sturgeons and paddlefishes)
Upper Cretaceous to Recent. Almost no internal ossification; platelike scales in isolated rows (Acipenseridae); snout enlarged and tactile (Polyodontidae); median fins chondrostean in having more fin rays than basal elements; tail heterocercal. Marine and freshwater, bottom suctorial feeders (sturgeons, Acipenseridae; Europe, Asia, North America) and plankton feeders (paddlefishes. Polyodontidae; China and North America). Length (sturgeons) up to 8.5 m (about 28 ft); weight to 1,400 kg (3,080 lbs.).

Order Polypteriformes (bichirs and reedfish)
Pleistocene to Recent. Typical chondrostean characters, such as ganoid scales. Dorsal fin modified into row of finlets; tail diphycercal; freshwater; Africa.

Critical appraisal. Because they are a fairly uniform group, the classification of the Chondrostei is difficult and unsettled. About 37 families are now recognized. The relationships of bichirs and the reedfish are especially controversial. Some authorities place them in a separate subclass; others conclude that they are related to the crossopterygians. (B.Sc./Ed.)

BOWFIN, GARS (HOLOSTEI)

The Holostei are one of the three major groups of rayfinned fish (Actinopterygii). Most paleontologists divide the Holostei into the Holosteans proper, which include living and extinct forms, and the Halecostomi, an extinct group.

The origin of the Holosteans and the Halecostomi is not fully understood, but it is believed that they arose from some advanced chondrostean fishes (the group that includes the sturgeon; see above). The Holostei were particularly abundant and diversified during the Mesozoic Era (65,000,000–225,000,000 years ago). Today they are represented by only two living genera, *Amia* (bowfin) and *Lepisosteus* (gar). One species of bowfin has been recognized, and about eight species of gar have been described so far.

Size range, distribution, and abundance

The gar occurs only in North America and Central America from southeastern Canada to Panama; it is not found west of the Rocky Mountains. The longnose gar (*L. osseus*) is the most widely distributed species. The gar is primarily a freshwater fish but sometimes ventures into saltwater or brackish water. The so-called alligator gar (*L. spatula*), one of the largest of freshwater fishes, is particularly abundant in the Everglades region of southern Florida, where it is caught locally as a food fish; it sometimes grows to a length of nearly three metres (10 feet) and may attain a weight of 136 kilograms (300 pounds). The names gar, garfish, and garpike are sometimes applied, especially in Europe, to the needlefishes (Belonidae), which are coastal fishes of warm seas and have very long and slender jaws. These fishes, however, are not closely related to the Holostei.

The bowfin, also known as grindle, mudfish, and dogfish, is found in sluggish waters from the Great Lakes to the Gulf of Mexico. It was once common throughout Europe but is now extinct there. Female bowfins reach a length of about 75 centimetres (30 inches) and weigh up to 3.5 kilograms (eight pounds); males are smaller. Bowfins eat all kinds of fish and invertebrates and are sometimes destructive to game-fish populations. Bowfins are seldom caught as food fish.

Drawing by J. Helmer, from (top centre) N.B. Marshall, *The Life of Fishes* (1965), Weidenfeld & Nicolson, London, (bottom) A.S. Romer, *Vertebrate Paleontology* (1966), University of Chicago Press

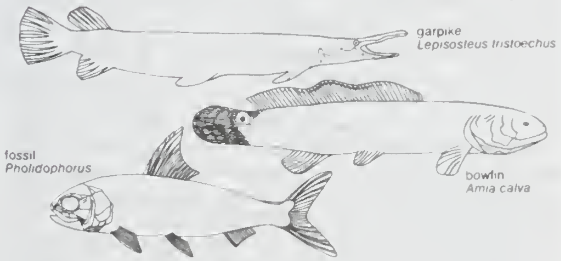


Figure 15: Living and fossil forms of Holostei.

Natural history. *Reproduction.* The bowfin spawns in weedy areas along the edges of streams and lakes. The male constructs the nest and guards the eggs as well as the newly hatched young. The young bowfin has an adhesive organ at the tip of its snout that enables it to cling to weeds. The fish grows rapidly and at the end of its first year may be as long as 23 centimetres (nine inches).

The female gar lays its large, yolk-filled eggs in shallow water in the spring. The gar hatchlings grow rapidly, feeding on minnows. The long rows of needle-sharp teeth are effective in capturing fast-swimming prey.

Ecology. Gars and bowfins are voracious predators, feeding on invertebrates and other fishes. All the amiiform fishes (the bowfin order) of former times were probably predaceous. For the most part they were a marine group; the modern bowfin, however, is confined to freshwater. Because of its highly developed air bladder, which can also function as a lung, the bowfin is able to live out of water for as long as 24 hours.

Gars are known to occasionally venture into saltwater, but apparently they do not attempt to feed there. They often float quietly at the surface of sluggish waters, breathing atmospheric air.

Form and function. *General features.* The Holostei are characterized by having the dermal bone of the upper jaw (maxilla) freed from the cheek elements and attached to the skull only in the ethmoid region or near the nasal chambers. The palate is separated from the cheek elements, and the adductor mandibulae muscle, which closes the jaws, is larger and more subdivided than it is in the chondrosteans. Primitively, the centrum (*i.e.*, the lower, heavy part of a vertebra) surrounding the notochord (a flexible rod that passes through the vertebral column) was absent, but this structure apparently developed independently in most of the holostean orders. The scales, too, were primitively rhomboidal, or diamond shaped, with a reduced or absent layer of dentine (*i.e.*, the substance of which teeth are largely made). The scales, however, became thin and cycloidal (*i.e.*, rounded and overlapping) in several groups. The fin rays of the unpaired fins are always equal in number to their basal supports, and the fins themselves may or may not be bordered at the anterior end by fulcra (*i.e.*, modified scales or spines). The caudal, or tail, fin is typically hemiheterocercal (*i.e.*, the upper lobe larger than the lower) and externally symmetrical. The braincase is always composed of separate ossifications (centres of bone formation) that resemble, in number and placement, those found in the teleosts.

The division Holosteans includes the orders Semionotiformes, Pycnodontiformes, Amiiformes, and perhaps Pachycormiformes. In these orders the preoperculum (an L-shaped bone anterior to the operculum, or gill cover) is tied to the palatal elements and provides part of the originating area for the adductor mandibulae muscle.

Extant groups. The order Seminonotiformes includes two families. The oldest known holostean, *Acentrophorus* (Upper Permian—about 225,000,000,000–250,000,000 years ago), belongs to the Semionotidae. Members of this family have small mouths and strong teeth; heavily ossified (*i.e.*, composed of true bone rather than cartilage) dermal bones; and hemiheterocercal tails. The body may be fusiform (*i.e.*, tapered at both ends), as in *Semionotus*, or flat and disk-shaped, as in *Dapedium*.

The other semionotiform family, the Lepisosteidae, includes the living gar. The characteristic snout of the gar is greatly lengthened by multiplication of the small tooth-bearing bones anterior to the eye. The premaxilla bone is situated at the anterior end of the series; the maxilla bone itself elongated seems to be reduced to a small, bony sliver at the angle of the mouth.

Body structure of living gars

The body of the gar is encased in an armour of thick, diamond-shaped, enamelled scales. The jaw ends in a beak that, in the alligator gar, is broad and relatively short; in the longnose gar the beak is long and forceps-like. The dorsal and anal fins, both located far back on the body, are without spines and have fewer than 12 rays each.

The Amiiformes, represented today by one species of bowfin, include about six families that show considerable diversity in the length of the jaw, the development of the teeth and fins, and details of the dermal skull pattern. In general, the earlier amiiforms had well-developed rhombic scales and a persistent notochord. In later forms the scales usually became thinner and cycloidal. Ossified centra developed around the notochord, either restricting it or eliminating it. The caudal fin was either forked or lobed. The amiiform body was generally fusiform, similar to that of the living bowfin.

The bowfin has a long, spineless dorsal fin with about 58 rays. This extends over most of the back to near the tail. The males have an orange- or yellow-encircled dark spot near the tail. In females either the outer circle or the entire marking is absent. Bony plates cover the head; the rest of the body has cycloid (*i.e.*, fan-shaped) scales.

Extinct groups. In some ways the Pachycormiformes superficially resemble certain living teleosts, such as mackerels and swordfishes. Their bodies are generally fusiform, with a widely forked caudal fin, a fairly wide gape, and moderately well-developed teeth. The pectoral fins may be elongated and the dorsal and anal fins somewhat enlarged.

The Pycnodontiformes, which may be related to the Semionotiformes, are unique among the holosteans in having their upper and lower dentitions modified to form an open pavement of crushing teeth. In many cases, however, the anterior teeth of the premaxilla and the dentary are incisiform and thus must have been used for grasping (as such teeth are in the living porgies and sparids). In addition to skull modifications related to feeding, the pycnodonts are characterized by deep, almost disk-shaped bodies, elongated anal and dorsal fins, and an externally symmetrical caudal fin. In a number of genera scales are absent on the posterior part of the body, a condition that apparently increased flexibility. Scales were usually present but modified on the anterior half. The body and fin form of pycnodonts suggest that they were fairly fast and powerful swimmers. The affinities of this order remain problematical, as the ossification pattern of the braincase and the caudal-fin skeleton do not closely resemble those of other Holosteans or Halecostomes.

The second major division of the subclass Holostei is the Halecostomes; all are relatively small, fusiform fishes. The group presently includes only one order, the Pholidophoriformes. Some authorities include a second order, Leptolepiformes.

Historical origin of gars and bowfins

Evolution. The gars probably arose in the Cretaceous Period (136,000,000 to 65,000,000 years ago) from some semionotid stock. They are known from freshwater Tertiary deposits in India, Africa, North America, and Europe. The bowfins also made their first appearance in Cretaceous times. Pycnodont fossils range from the Upper Triassic to the Eocene (from about 190,000,000 to 50,000,000 years ago). Pachycormiforms are known only from marine rocks of Jurassic and Cretaceous age. The pholidophorids are known from the Triassic to the Cretaceous; the other pholidophoriform families all ranged from the Jurassic to the Cretaceous.

Among the seven families presently assigned to the Pholidophoriformes, the pholidophorids probably show the closest resemblance to the early teleosts. Trends toward thinning of the scales and the loss of ganoin (an enamel-like material) on the fin rays, along with the dermal-bone pattern and the development of intermuscular bones, point toward the teleosts. The major difference between the pholidophorids and the teleosts is in the structure of the caudal-fin skeleton. In pholidophorids of the early Jurassic, the caudal fin was still structurally heterocercal, with a fairly stiff axial lobe. Modification toward the teleost condition involved changes that brought about equal flexibility of the upper and lower lobes. The other six families currently assigned to the order Pholidophoriformes are specialized in various ways, but none can be regarded as involved in the ancestry of the teleosts.

Classification. Groups marked with a dagger (†) are extinct and known only from fossils.

Distinguishing taxonomic features. The principal features on which classification of the Holostei is based include general body shape, scale structure, and the number and placement of head bones.

Annotated classification.

SUBCLASS HOLOSTEI

Tail hemiheterocercal; maxilla free of preopercle; rays of median fins approximately equal in number to basal elements; trend toward thinning of scales and loss of ganoid (enamel) layer.

Division Holosteans

Preopercle intimately bound to and supporting the posterior border of the palate.

Order Amiiformes (bowfin and fossil relatives)

Lower Triassic (about 210,000,000–225,000,000 years ago) to Recent; body generally fusiform; early forms with well-developed rhomic (diamond-shaped) scales and persistent notochord; scales thinner and cycloidal (fan-shaped) in later forms; 6 families; 1 living species.

†*Order Pachycormiformes*

Lower Jurassic (about 160,000,000–190,000,000 years ago) to Upper Cretaceous (about 65,000,000–100,000,000 years ago); body fusiform, caudal fin widely forked, long snout; 2 families; Europe and North America.

Order Semionotiformes

Upper Permian (about 225,000,000–250,000,000 years ago) to Recent; 2 families of widely divergent fishes; fossil Lepidotidae with normal holostean fusiform bodies, which became relatively deep and slab sided in some members; marine and freshwater, widely distributed; gars (Lepisosteidae) are elongated, long snouted, primarily freshwater predators, extant in North America.

†*Order Pycnodontiformes*

Upper Triassic (about 190,000,000–210,000,000 years ago) to Eocene (about 38,000,000–54,000,000 years ago); upper and lower teeth modified to form crushing pavement; body nearly disk-shaped; anal and dorsal fins elongated; caudal fin externally symmetrical.

Division Halecostomi (or Halecostomes)

Relatively small; body fusiform; preopercle not buttressing the bones of the palate.

†*Order Pholidophoriformes*

Middle Triassic (about 200,000,000 years ago) to Lower Cretaceous (100,000,000–136,000,000 years ago); holosteans with some trends toward teleosts, notably: loss of ganoin from fin rays, scales, and dermal bones; loss of peg and socket joints between scales; about 7 families; marine and freshwater, of wide distribution.

Critical appraisal. According to some authorities, the Leptolepiformes, a teleost group, should be included among the Halecostomes. This opinion indicates that the boundary between the Halecostomes and the Teleostei is difficult to define. The family Pholidophoridae, in particular, has a skull pattern almost identical with that of the leptolepids; the feeding mechanisms are also quite similar. (Ed.)

The modern bony fishes, or teleosts (Teleostei)

General characteristics of teleosts

The name teleost is applied to members of the Teleostei, a diverse group (infraclass) that includes virtually all of the world's important sport and commercial fishes, as well as a much larger number of lesser known species. The infraclass is distinguished primarily by the presence of a homocercal tail; *i.e.*, one in which the upper and lower halves are about equal. The teleosts, sometimes called "advanced bony fishes," comprise some 20,000 species (about equal to all other vertebrate groups combined), with new species being discovered each year.

The great abundance of some large species, such as the tunas and halibuts, and of smaller species, such as the various herrings, have made teleost fishes extremely important to mankind as a food supply. In almost every part of the world local fishes are used as food by men at all stages of economic development. In addition to being a commercial food resource, teleost fishes provide enjoyment to millions of people and in many countries of the world support a large sport fishing industry. As aquarium subjects both marine and, especially, small freshwater teleosts provide esthetic beauty for millions of aquarists, supporting a multimillion dollar industry. Part of the interest in teleosts as aquarium subjects is derived from the great diversity of their anatomical structures, functions, and colour. Indeed, these fishes vary more in structure and behaviour than do all the mammals, birds, reptiles, and amphibians combined. Teleosts range in size from tiny gobies less than an inch long when fully adult to large marlins exceeding 3.4 metres (11 feet) in length and 550 kilograms (1,200 pounds) in weight. Another large fish, the ocean sunfish (*Mola*), reaches at least three metres (10 feet) and may weigh more than 900 kilograms (2,000 pounds).

Natural history

Different species and groups have widely varying life cycles and behaviours. Teleost fishes are adapted to widely varied habitats from cold Arctic and Antarctic oceans that remain colder than the freezing point of fresh water to desert hot springs that reach temperatures over 38° C (100° F). Usually a particular species will have a restricted temperature tolerance, often between 10° to 20° F above and below the mean temperature of its environment. In terms of physical habitat, the advanced bony fishes are adapted to a wide and interesting variety of conditions, from fast, rock-laden torrential streams in the Himalaya Mountains to the lightless depths of ocean trenches 10,670 metres (35,000 feet) below the surface, where many species manufacture their own light.

Reproductive behaviour

A species of teleost fish is usually restricted to one kind of habitat at any given stage of its life cycle. It may occupy this habitat throughout its life, or it may change habitats as it grows older. The various species living in rocky marine shores, mud flats, sandy shores, and coral reefs are usually all different, a diversity of habitat adaptation reflected in the life cycle, behaviour, locomotion, anatomy, and reproduction. Some cyprinodonts or killifishes, for example, confined to annual ponds in Africa and South America, live only the few months during the rainy season that their ponds retain water. They hatch from eggs buried in the mud, grow up, lay their own eggs in the mud in the short space of four to eight months. Such fishes are known as annuals. Others, such as the Pacific salmon, hatch from eggs laid in the gravel of cool, temperate zone streams, spend their first year growing in the streams, then enter the sea to grow and migrate for two, three, or four years, finally returning to the streams where they first grew up.

There they lay their eggs and die. Such fishes, spawning in fresh water and living most of their lives in the ocean, are called anadromous.

Many freshwater and marine teleosts lay eggs on rocks or aquatic plants, the male and sometimes the female defending the eggs and even the young against predators. Many of these fishes will live two, three, or four years or more, usually spawning in the spring in temperate regions and in the rainy season in the tropics. There are large numbers of teleosts, especially those in the ocean, whose breeding habits are unknown. Most fishes lay numerous eggs, often simply scattering them over plants or in the open ocean, where they provide food for many organisms, only a few young surviving to adulthood. Most offshore marine teleosts lay planktonic (free-floating) eggs, whereas most freshwater fishes lay demersal eggs; *i.e.*, eggs that sink to the bottom. Some teleosts, such as certain of the perchlike African cichlids, some catfishes, and some marine fishes (*e.g.*, cardinal fishes) are oral brooders, the male or female incubating the eggs in its mouth.

Some fishes, for example, some of the sea perches (Serranidae), are functional hermaphrodites, one individual producing both sperm and eggs. Self-fertilization is evidently possible in some cases, but more often the fish plays the male and female roles alternately. In some species an individual is a male during the early part of its adult life and a female later.

About a dozen families of teleosts produce living young. In some the eggs are abundantly supplied with yolk and merely hatch in the ovary (ovoviviparous), in others the eggs have little yolk, the young hatch at a relatively undeveloped state and are nourished by a placenta-like structure of the ovary (viviparous).

The behaviour of teleosts is as varied as their other attributes. Some oceanic fishes travel in close-ordered schools, seemingly responding to predation from larger fishes almost as a single organism. Larger, predatory fishes are usually solitary and hunt or wait for prey alone. Many marine shore and freshwater fishes establish territories during the breeding season, and some may travel in relatively loose schools or shoals when not breeding. Many kinds of teleosts enter into symbiotic relationships with other species of fishes and organisms, for example, a small, blind goby along the California coast lives together with a shrimp in the shrimp's cavelike tubular dwelling. The shrimp carries food to the goby while the goby keeps the shrimp's burrow clean. Many species of wrasse pick parasites from larger fishes, even entering the mouths of these fishes to clean the gill chambers. On the other hand, in South America some small catfishes, 2.5 to 10 centimetres (one to four inches) long, appear to be parasitic on certain other species of catfishes that reach lengths of over 2.4 metres (eight feet). The smaller fishes enter the mouths of the larger catfishes and feed on gill tissue. Unfortunately, little is known about the behaviour of most teleosts, but study of the details of their behaviour has been greatly increased in the last several decades.

Locomotion

Defining teleost fishes by functional morphology is hard, for they have evolved into many diverse shapes; but if a relatively simple teleost, a trout, is examined, its basic swimming motion can be determined. Forward motion is provided by bending of the body and caudal fin; waves of muscular action pass from the head to the tail, pushing the sides of the body and tail against the water and forcing the fish forward. The structure of the tail and the efficiency of the swimming mechanism is the prime character that distinguishes teleosts from other, "lower," fishes. The dorsal fin and the anal fin (a ventral median fin) are used partly to aid in stability and in turning, and partly in forward locomotion. The paired pelvic or ventral fins and the paired pectoral fins behind the head are used to help stabilize the body and to turn the fish. The fusiform shape of the trout reduces the turbulence and drag of water flowing over the fish's body, offering least resistance to the water.

The head of the fish must be adapted for feeding, breathing, and detecting prey and enemies. At the same time it must be relatively streamlined, offering as little resistance to the water as possible. The head of a trout is well formed for these functions by being fusiform but expandable, where necessary, to take in food and water. The fish forces the water in one direction, into the mouth, over the gills, and out the gill slits. Back flow is prevented by valves at the mouth and by the gill covers. The fish, however, can eject undesirable particles and water out the mouth by special action. The teleost head is efficient in having eyes and organs for the sense of smell located in optimum spots for seeing and smelling food. At the same time these organs offer little resistance to water flowing over the head.

For a complete classification of teleosts, see the *Annotated Classification* in the first section of this article; additional taxonomic information is presented in the following sections on the major groups of teleosts.

(S.H.W.)

MAJOR TELEOST GROUPS

Bonefish, tarpons, ladyfishes, and allies (Elopiformes)

The order Elopiformes contains about 12 species of marine and brackish water fishes, the best known of which are bonefish, tarpons, and ladyfishes. Most taxonomists recognize two living suborders of elopiforms: Elopoidei, which consists of two living families; and Albuloidei, which contains one living and one extinct family. A few elopiforms are prized gamefishes, but only the Pacific tarpon (or oxeye) is of economic importance as food, supporting a major fishery in Southeast Asia. The terms "ladyfish" and "bonefish" have both been used for *Elops saurus* and *Albula vulpes.* In this article the name ladyfish is applied only to *Elops* and bonefish only to *Albula.* The tarpons and ladyfishes are fast-swimming predators with adult lengths of up to 2.5 metres (approximately eight feet) in tarpons and about one metre (three feet, three inches) in ladyfishes. The bonefish and Japanese gisu are specialized bottom feeders. All except the gisu are coastal fishes of warm oceans, most common in latitudes from 20° N to 20° S.

Primitive position among teleosts

Elopiforms are of interest to the ichthyologist as the most primitive living teleost fishes, standing in much the same relation to the higher bony fishes as do the egg-laying mammals (monotremes) to other mammals. As is usual with primitive groups, the elopiforms have an extensive fossil record, with many more fossil than recent species.

GENERAL FEATURES

Despite their archaic structure, elopiforms are related by their life history to more specialized groups such as the eels. Like eels, elopiforms have a ribbonlike, translucent, pelagic larva (leptocephalus) that undergoes a striking metamorphosis involving shrinkage to about half the maximum larval size. The bonefish, tarpons, and ladyfish spawn close to shore. The eggs are shed and fertilized in shoal water, sinking to the bottom. Elopiforms are prolific breeders; a large Atlantic tarpon (*Tarpon atlanticus*) was estimated to contain more than 12,000,000 eggs, about seven times as many as in the proverbially fecund cod. The newly hatched leptocephali may be carried out to sea by offshore currents, but metamorphosis only occurs close inshore, and it is probable that larvae carried far out to sea die.

During or immediately after their metamorphosis, the postlarvae migrate inland and accumulate in brackish pools or creeks, often connected with open water only at extreme high tide. Such environments are stagnant and low in oxygen, and air breathing (see below) is an important aid to survival. The juvenile fish feed on small crustaceans, insect larvae, and other small animals, moving back to the sea as young adults. The gisu (*Pterothrissus gissu*), which differs from other elopiforms in inhabiting deeper and colder waters, is apparently an exception to this pattern of development, larval life and metamorphosis

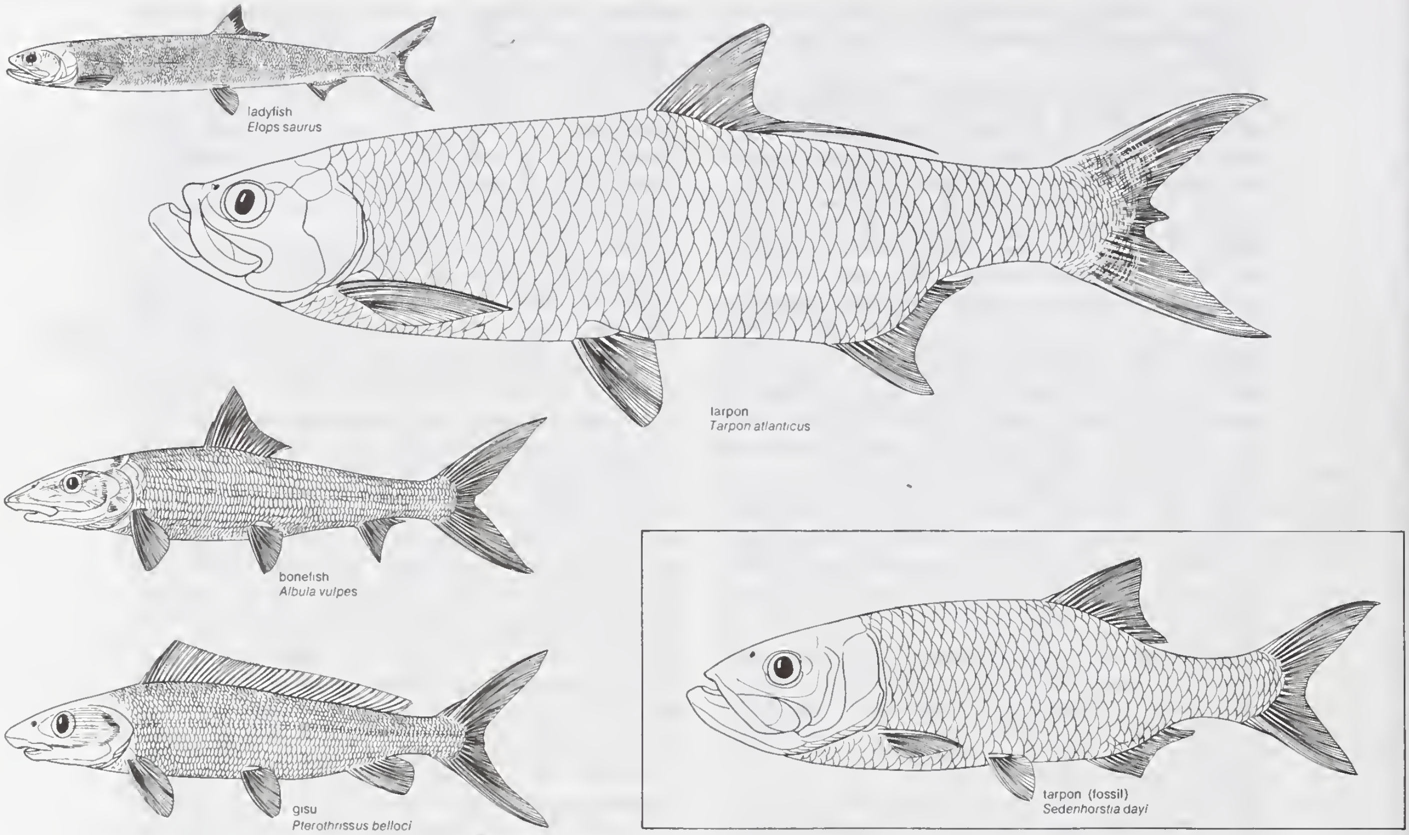


Figure 16: Body plans of several elopiform fishes.

From (*Elops saurus, Tarpon atlanticus, Albula vulpes*) *McClane's Standard Fishing Encyclopedia* edited by A J McClane illustrated by Richard E Younger, Copyright © 1965 by Holt, Rinehart and Winston, Inc , reproduced by permission of Holt, Rinehart and Winston, Inc , (*Pterothrissus belloci*) from M Poll, "Resultats scientifiques de l Expedition oceanographique belge dans les eaux cotieres africaines de l'Atlantique Sud (1948–1949)." III Teleosteens malacopterygiens, Vol IV, fasc 2, 1953, p 14 fig 5

taking place in deep water, without the postmetamorphic inland stage that is experienced by other elopiforms.

With the exception of the gisu, elopiforms are coastal fishes, able as adults to enter brackish or fresh water. Adult ladyfishes and tarpons are typical predators of coastal waters, feeding mainly on other fishes. The Atlantic tarpon is renowned for leaping out of the water; the Pacific tarpon (*Megalops cypinoides*) and ladyfishes (several species of *Elops*) behave similarly, "rolling" at the surface. The purpose of this behaviour seems to be the intake of air. Like all of the other primitive teleosts, the elopiforms possess an open duct to the swim bladder, and air that is taken in at the mouth can be passed into the swim bladder.

Behaviour related to air intake

In tarpons the swim bladder is lunglike, partially compartmented and highly vascularized. Tarpon are obligate air breathers, dying from asphyxiation if prevented from reaching the surface, an unusual condition for a species in which adults normally inhabit well oxygenated waters. Such an adaptation, however, is certainly advantageous in the stagnant pools where postlarval life is spent. The tarpons exhibit a further modification of the swim bladder, a pair of forward outgrowths that contact the auditory region of the braincase and are partially enclosed in bony bullae, a modification that presumably improves the sense of hearing.

The bonefish is a bottom feeder in shallow water along coastal areas, coming in with the tide and grubbing with its snout for worms and shellfish, which it is able to crush with its rounded palatal teeth. It can expose prey buried in the sand by directing a jet of water from its mouth. The deep-water gisu feeds mainly on worms, probably by similar methods.

PALEONTOLOGY AND CLASSIFICATION

Fossil history. The family Elopidae is the only extant teleostean family whose fossil record extends back into the Jurassic (about 140,000,000 years ago). The Upper Jurassic genus *Anaethalion* is included in this family on the basis of some forms that were extremely similar to the modern *Elops*. The genera *Notelops*, from the Lower Cretaceous of Brazil, and *Osmeroides*, widely distributed in Upper Cretaceous seas, were probably true elopids, but the allocation of numerous little-known Cretaceous genera currently placed in the Elopidae, often on the basis of negative evidence, must be considered tentative. The earliest known member of the tarpon family appears to be the fossil *Sedenhorstia*, from the Upper Cretaceous of Europe and Lebanon. Fossils assigned to *Megalops* appear in Eocene deposits. The earliest member of the extinct suborder Pachyrhizodontoidei, *Rhacolepis* (from the Lower Cretaceous of Brazil), was small and resembled the ladyfishes, but later, Upper Cretaceous, members of this group become considerably more specialized. *Pachyrhizodus*, from the Cretaceous chalks of Europe and North America, exceeded three metres in length and superficially resembled a tuna. Pachyrhizodontoids may well have been large fast-swimming predators of the open sea, a niche which is now filled by the tunas.

The genus *Albula* has been recorded in the Upper Cretaceous of North America and in the Paleocen of Europe, but typically albulid tooth plates occur in Lower Cretaceous deposits. Several Cretaceous and Tertiary albulid genera have been described on the basis of otoliths ("earstones"). *Istieus*, from the Upper Cretaceous of Germany, had characteristics that were very similar to those of the modern *Pterothrissus*.

Annotated classification.

ORDER ELOPIFORMES

Elopomorph fishes having a long but not eellike body; scales containing bone-cells; teeth on the parasphenoid bone; a transverse sensory canal across the snout; broad infra-orbital bones extending back to the operculum, 55–90 vertebrae, a primitive tail skeleton with 2 free vertebrae supporting 6–7 hypural bones, tail fin large and forked.

Suborder Elopoidei

Mouth terminal and snout unmodified; 2 supramaxillaries; many branchiostegal rays (23–35); teeth small; large gular plate between the lower jaws; 7 hypural bones.

Family Elopidae (ladyfish or tenpounder)

Upper Jurassic to Recent. Very generalized fish, the living

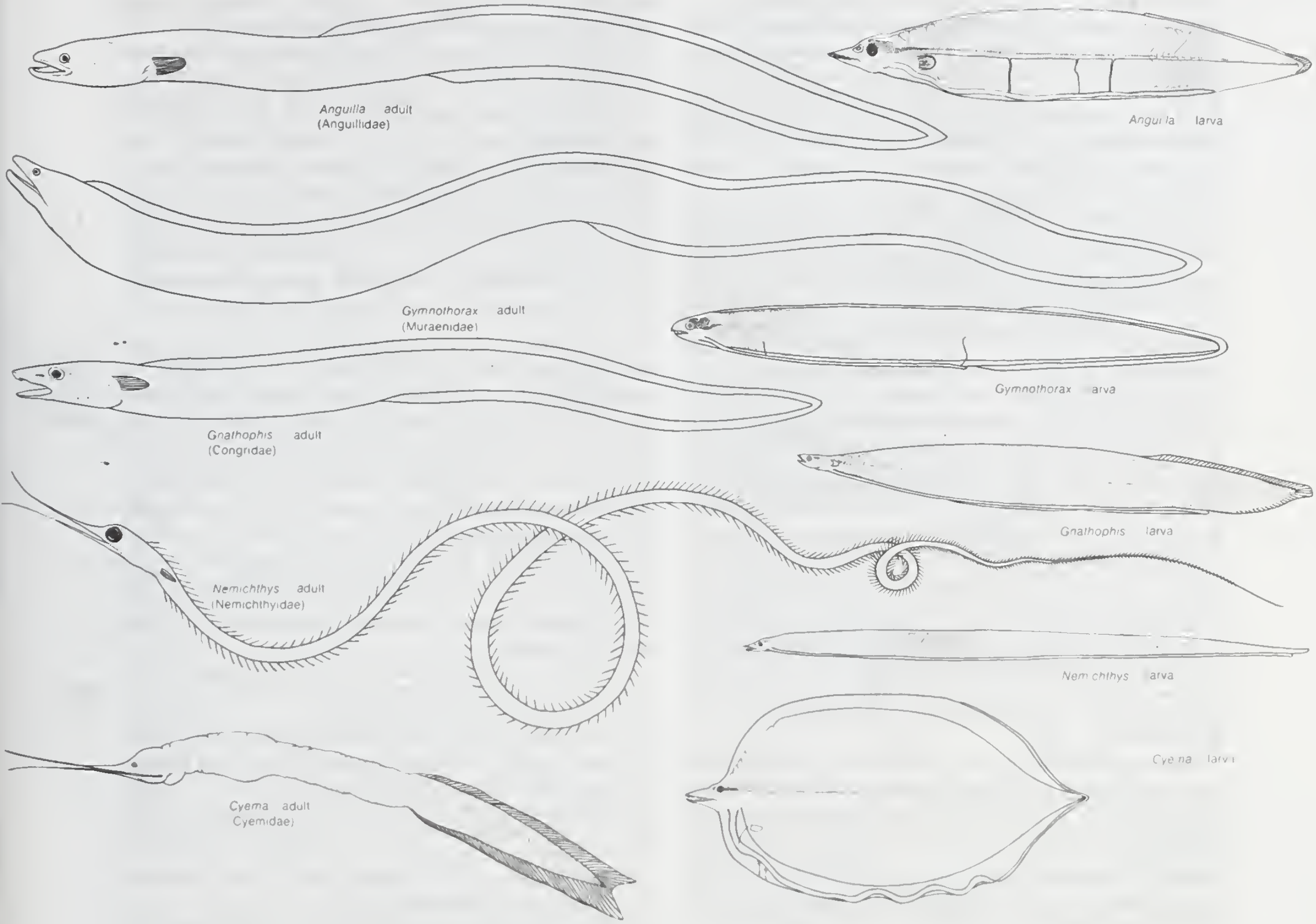


Figure 17: Adults and larvae of representative eels.

Drawing by J Helmer based on (Cyema adult) by courtesy of Carl L Hubbs from (*Gymnothorax* larva) E Herald *Living Fishes of the World*, Doubleday & Co Inc (*Anguilla* adult *Nemichthys* adult) P H J Castle World of Eels (1968) in *Tuatara* vol 16 no 2 (*Anguilla* larva *Gnathophis* adult *Gnathophis* larva) *Zoology Publications from Victoria University of Wellington* (1963) (others) *Transactions of the Royal Society of New Zealand* (1965)

forms having 32–35 branchiostegal rays and the swim bladder unmodified. Length to 0.9 m (about 3 ft); weight to about 13 kg (28½ lb). One living genus, *Elops*, with 5 or 6 species, circumtropical. Numerous fossil genera.

Family Megalopidae (tarpons)
Swim bladder partially cellular, lunglike, and connected with the ear; scales large; 23–25 branchiostegal rays. Two living genera, *Megalops* and *Tarpon*, each with one species, recently shown to differ in the ear to swim-bladder connection. *Tarpon* confined to the Atlantic; *Megalops* to the Indian and Pacific oceans, ranging from Africa to the Philippines. *Tarpon* length to 2.5 m (about 8 feet), weight to 150 kg (about 330 lb); *Megalops* to 1.5 m (5 ft).

†Suborder Pachyrhizodontoidei
Fossil; Lower to Upper Cretaceous. Parietal bones separated by supra-occipital; gular plate absent; teeth large, pointed, and set in deep sockets. Length to 3 m (10 ft).

Suborder Albuloidei
Snout enlarged; mouth small and underslung; crushing teeth on palate; single supramaxillary bone; gular plate small or absent; 6 hypural bones.

Family Albulidae (bonefishes and gisu)
Upper Cretaceous to Recent. Characteristics those of suborder. Two living genera: *Albula* (2 species) and *Pterothrissus* (2 species). Length to 70 cm (28 in.), weight to about 6.5 kg (14½ lb). *Albula* inhabits coastal waters, *Pterothrissus* deeper waters (below 200 m [650 ft]).

†Family Phyllodontidae
Fossil. Upper Cretaceous and Early Tertiary. Known only from thick tooth plates, with many superimposed series of replacement teeth. Five genera: from Africa, Europe, and North America.

Critical appraisal. The chief problem in defining and classifying the elopiforms is the reliance on primitive features, which are merely negative evidence against relationships with other, more advanced groups. Similar problems arise among the most primitive members of all animal groups. The possession of a leptocephalus larva places the elopiforms as the basal order of the cohort Elopomorpha, which also contains the eels, halosaurs, and notacanths. But it is only on negative evidence (absence of specializations) that the elopoids and albuloids are associated, and the absence of any notable specializations in elopoids makes the placing of fossil forms in this group rather doubtful. (C.P.)

Eels (Anguilliformes)

True eels are elongated or even wormlike bony fishes of the order Anguilliformes and include the common freshwater eels as well as the voracious marine morays. Only the freshwater eels (family Anguillidae), which are in places abundant and greatly valued as food, are of major economic importance. Regardless of their final habitat, all eels probably pass through an extended larval phase (leptocephalus) in the open ocean. They undergo metamorphosis to a juvenile stage that is a smaller version of the adult. At maturity, eels range from 10 centimetres (four inches) (in the deep sea *Cyema atrum*) to 3.5 metres (11½ feet) (in the moray *Thyrsoidea macrura*). They occur to considerable depths in most oceans and are greatly diverse in tropical seas. They range in colour from drab gray or black (in deep sea species) to colourful and patterned (in tropical reef species).

NATURAL HISTORY

Eels have a remarkable life cycle. Broadly, it consists of development and early growth in the open ocean: the planktonic (free-floating) dispersal of eggs and larvae, Life cycle metamorphosis, juvenile and adult growth, and the migration of maturing adults to an oceanic spawning area. Eels share the leptocephalus phase with several other orders (Elopiformes, Saccopharyngiformes, and Notacanthiformes). A prolarva, hatching from a relatively large egg (up to 2.5 millimetres [about $^{3}/_{32}$ inch] diameter), rapidly becomes a leaflike leptocephalus, which floats in the surface layers of the open ocean for as long as $2^{1}/_{2}$ years before metamorphosing.

Although the leptocephali were once thought to have been fishes of a distinct group, their relationship with the Anguilliformes was soon recognized from transitional specimens that showed larval and adult characters. They proved so difficult to identify, however, that new larval types were named as species of the genus *Leptocephalus* (though they cannot actually be considered different species from the adults that produced them), accounting for the several hundred forms known.

Leptocephali are not uncommon in the upper 500 metres (roughly 1,600 feet) of the ocean, a distribution that may be associated with the availability of food (diatoms and minute crustaceans). Their predators include various pelagic fishes. In tropical eels larval life is possibly four to six months, but temperate species may spend upwards of a year as larvae. During this time leptocephali, in the presence of suitable currents, may disperse widely from the adult spawning area. Working on massive collections of larvae from 1905 to 1930, a Danish biologist, Johannes Schmidt, established the early life history of the European and American freshwater eels. Although parts of his work have been questioned, his description of a western Atlantic spawning and a trans-Atlantic dispersal of leptocephali of these eels still stands.

After reaching full growth, the larva begins a rapid metamorphosis in which the following progressive changes are typically involved: the body becomes cylindrical and greatly reduced in bulk, perhaps by as much as 90 percent by weight; the anal vent advances from its subterminal position to about the midpoint; larval teeth are lost; the snout becomes rounded; the dorsal fin originates farther forward; and the larval melanophores (black pigment cells) disappear. Other changes, such as the loss of the pectoral fins or a reduction of body length, may also occur.

By courtesy of Direction generale des Peches maritimes, Ministere de l'Industrie et du Commerce, Quebec

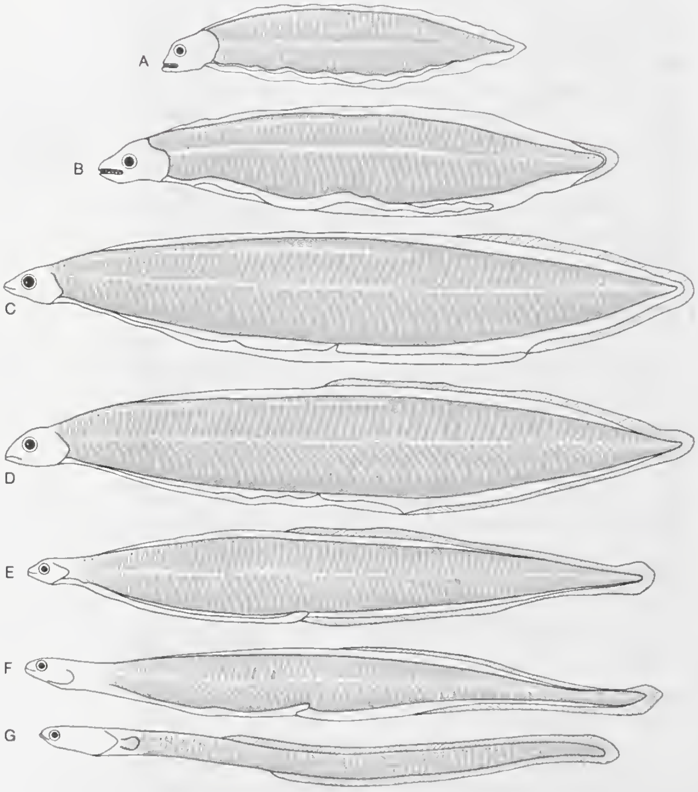


Figure 18: *Metamorphosis of the American eel.*
(A–C) Larvae or leptocephali of various sizes. (D–F) Larvae in the process of metamorphosis. (G) Glass eel, mature stage.

Leptocephali are markedly unlike their adults and the metamorphic changes are so great that a fundamental problem arises in the correlation of the great variety of known leptocephali with their adults. Metamorphosis has been followed through in aquariums and deduced from progressive growth series in plankton samples. Certain characters survive metamorphosis and are important in the recognition of eel species. These include the number of muscle segments (myomeres); the development of dorsal, anal, and caudal fin rays; and the relative positions of the renal vessels and gallbladder. In many leptocephali, the larval melanophores also remain into the juvenile (or elver).

The process of metamorphosis

Metamorphosis involves physiological and behavioral as well as structural changes, particularly those related to the assumption of a deep-sea, shallow-water, or freshwater mode of life. Metamorphosis is the mechanism by which the leptocephalus, after a period of growing, feeding, and competing with other similarly organized planktonic animals, can enter a markedly different habitat, where body shape, differentiated feeding mechanisms, sense organs, and body coloration, play an important role in survival. Metamorphosis in all eels is probably completed in the open ocean. The annual invasion of freshwaters by *Anguilla* elvers is a locally well-known process; it occurs during October–March in Europe and in spring in other temperate regions.

During several years' growth to maturity, eels are essentially carnivores, feeding diversely on planktonic or benthic (bottom-living) animals. Maturity is reached after about 10 years in the European freshwater eel but possibly much earlier in tropical marine species. The process of growth and maturation has been most closely studied in the European freshwater eel. In this species both sexes pass through successive phases of neutrality, precocious feminization, and juvenile hermaphroditism prior to becoming definitively male or female, the sex being mainly determined by environmental factors.

All eels apparently undergo a short or long distance migration at maturity to a spawning area within the area of adult distribution (in most tropical marine eels) or some distance from it (temperate Anguillidae and Congridae). These areas are generally located over the continental slope or in ocean basins some distance offshore.

During their juvenile and adult life, most eels are solitary fishes, swimming slowly by means of sinuous lateral movements of the body and median fins. Some species burrow rapidly, using a pointed tail and backward body movements. Morays and congers inhabit rock crevices, while certain congrids (Heterocongrinae) form vast colonies of several hundred individuals in tropical reef areas.

FORM AND FUNCTION

An eel is distinguished externally from most other fishes by its elongated body, which is seldom laterally compressed. A continuous dorsal, anal, and caudal fin runs around the tail tip; pelvic fins are always absent; and gill openings are usually reduced. The body covering is usually scaleless. Minor departures from this overall body plan occur in the various eel families and are correlated well with different modes of life.

Typically, a leptocephalus is elongate, laterally compressed, transparent, and gelatinous, with prominent, W-shaped myomeres and sharp, forwardly directed larval teeth. At full growth, eel larvae are five to 10 centimetres (roughly two to four inches) but may be much larger (45 centimetres [about 18 inches] in the case of the snipe eel *Nemichthys scolopaceus*). Leptocephali are at least as diverse morphologically as their adults: some are filamentous, while others are deep-bodied, even resembling a small dinner plate in shape and size.

Pectoral and median fins are present in most leptocephali but may disappear during metamorphosis. Eyes are usually normal in shape but are occasionally telescopic, and the rostrum (an anterior beaklike projection) may be greatly extended forward from the snout. The attenuate viscera are located along the ventral aspect, below the myomeres.

There is usually a straight, unmodified intestine, and the anal vent is often forward of the tail. In some families the larval gut is swollen or festooned at various points along its length, a modification of unknown significance. A long liver, with a well-defined gallbladder, occurs anteriorly on either side of the intestine. The developing adult kidney lies at about midlength. The organs are supplied and drained by many vertical blood vessels, of which the last one or two are the largest. The position of the last of these vessels equates well with the division between precaudal and caudal vertebrae and indicates the approximate position of the adult kidney.

In most leptocephali melanophores occur almost anywhere on or in the body. These vary from minute and compact to large and irregularly shaped, often distributed along the gut and laterally in a variety of patterns.

In adults there is a strong tendency towards a reduction or loss of fins and a streamlining of the profile, sometimes also with an attenuation of the body. The gill region is variously elongated by a posterior displacement of the gill arches, accompanied by separation of the pectoral girdle from the cranium. There are well-developed pharyngeal tooth plates, and the pharynx has assumed an important function in the movement of food into the esophagus. It also serves as an effective pump in the virtual absence of suction by the gill covers (opercula) for the passage of the respiratory current. The opercular series is much reduced, although numerous curved branchiostegal rays (internal gill supports) are strongly developed to support the long throat wall. Respiration through the skin is important in *Anguilla* and probably also in other eels.

EVOLUTION, PALEONTOLOGY, AND CLASSIFICATION

Studies of the few known fossil eels and the comparative anatomy of adults and larvae suggest that eels arose in the Cretaceous Period from two or more types that had at least some characters of the Elopiformes (tarpons, bonefishes, etc.).

Distinguishing taxonomic features. The important characters in determining the taxonomic ranking of eels include: the number and arrangement of myomeres and fin rays; the relative positions of mouth, nostrils, gill openings, renal vessels, and gallbladder; and the number and arrangement of teeth, skull bones, and vertebrae. Also important in identification are larval structures and pigmentation patterns.

Annotated classification. The classification used in this article is a synthesis of the work of many authorities. It is certain to be further modified as more is learned about some of the little-known families.

ORDER ANGUILLIFORMES

Elongated, bony fishes of streamlined profile; continuous median fin with up to 650 soft rays; pelvic fins absent; reduced cycloid scales (*i.e.*, with skin-covered bone overlying a layer of fibrous connective tissue) sometimes present; swimbladder with duct; cranial bones reduced, particularly the palatoquadrate and opercular series; pectoral girdle detached from cranium; vertebrae up to 700, with fused centra and arches; a leptocephalus larva. There are 19 families, about 140 genera, and more than 500 species. One family in freshwater; the remainder marine, in all oceans, mainly tropical Atlantic and Indo-Pacific, to considerable depths.

Suborder Anguilloidei

Frontal bones of skull paired, supraoccipital present.

Family Anguillidae (freshwater eels)

Worldwide, but not on Pacific coast of America and south Atlantic coasts. Scales present, gill slits ventrolateral. Important as food.

Family Heterenchelyidae

Tropical Atlantic. No fins, mouth large.

Family Moringuidae (worm eels)

Tropical Indo-Pacific and western Atlantic. Anus in posterior half of body, degenerate, burrowing.

Family Xenocongridae (false morays)

Pantropical. Burrowing.

Family Muraenidae (morays)

Pantropical to subtropical. No pectorals, large mouth, often brightly coloured, voracious, sedentary.

Family Myrocongridae

South Atlantic. Laterally compressed, poorly known.

Suborder Nemichthyoidei

Frontals paired or fused, supraoccipital present or absent, paired nostrils close in front of eye.

Family Nemichthyidae (snipe eels)

Bathypelagic (deepwater), worldwide. Jaws greatly extended, minute teeth.

Family Serrivomeridae (sawtooth snipe eels)

Bathypelagic, worldwide. Jaws moderately extended; bladelike teeth on vomer bones.

Family Cyemidae (bobtail snipe eels)

Bathypelagic, worldwide. Jaws extended, body truncated.

Suborder Congroidei

Frontals fused, supraoccipital present.

Family Congridae (congers)

All oceans to considerable depths.

Family Muraenesocidae (conger pikes)

Pantropical. Large teeth, voracious.

Family Nettastomatidae (witch eels)

Deepwater. No pectoral fins.

Family Nessorhamphidae (duckbilled eels)

Bathypelagic. Duckbill snout.

Family Derichthyidae (neck eels)

Bathypelagic. Short snout.

Family Ophichthidae (snake eels)

All oceans, many branchiostegals, caudal reduced or absent.

Family Macrocephenchelyidae

Deepwater, Pacific. Rare.

Suborder Synaphobranchoidei

Frontals fused, supraoccipital present, 3rd hypobranchial directed backward, larva telescopic-eyed.

Family Synaphobranchidae

Deepwater, worldwide. Gill slits ventrolateral to ventral, united. Scales present.

Family Simenchelyidae (parasitic eels)

Deepwater, worldwide. Semiparasitic, mouth transverse, scales present.

Family Dysommidae

Deepwater, worldwide. Anus near pectoral region. No scales.

Critical appraisal. The eels were early split into the Colocephali (morays) and Enchelycephali (others). Later the suborder Carenchelyi was added (for the Derichthyidae). Two fundamental but undesignated groups, separated on the poorly evaluated character of fused versus separate frontal bones, have been considered the main evolutionary lines, forming the suborders Anguilloidei and Congroidei. In the absence of a workable alternative this division must stand. In 1940 L.S. Berg added the Nemichthyoidei, the subordinal position of which requires further investigation. Various osteological and larval characters suggest that the Synaphobranchoidei, as recently recognized, may truly represent a further lineage. Some would include still another suborder, the Saccopharyngoidei. These strange deep-sea fishes, while closely related to true eels, appear distinctive enough to merit separate ordinal rank as the Saccopharyngiformes.

The limits of the 19 living eel families are now broadly defined. However, their interrelationships and the existence of subordinal groupings will only be firmly established by an overall survey of a range of characters and not only of specific structures throughout the order. Rare forms need investigations and larval characters require consideration. Closer study is needed of environmental effects on segmental features, and fundamental questions of species nomenclature have yet to be settled. (P.H.J.C.)

Herrings, anchovies, and allies (Clupeiformes)

The order Clupeiformes, containing some of the world's most numerous and economically important fishes, includes more than 400 species, about 20 of which provide more than one third of the world fish catch. They are by far the most heavily exploited of all fish groups. Most

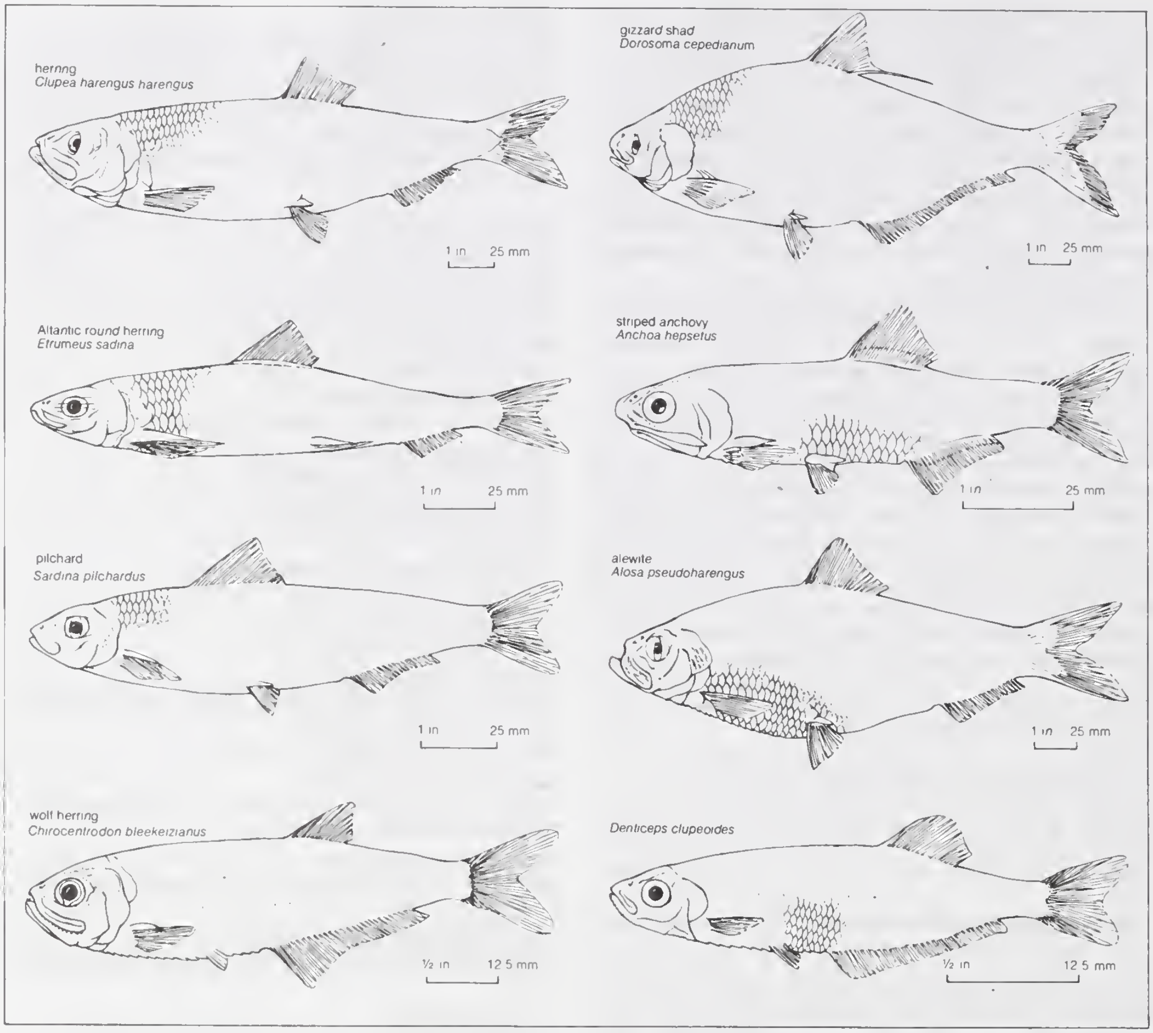


Figure 19: Body plans of representative clupeiform fishes.

Drawing by A. Murawski based on (*Clupea harengus, Etrumeus sadina, Anchoa hepsetus, Alosa pseudoharengus*) A.H. Leim and W.B. Scott, *Fishes of the Atlantic Coast of Canada* (1966), Fisheries Research Board of Canada reproduced by permission of Information Canada, (*Denticeps clupeoides, Chirocentrodon bleekerzianus*) *Bulletin of the American Museum of Natural History* (1966), (*Dorosoma cepedianum*) D.S. Jordan, *A Guide to the Study of Fishes*

clupeiforms are small marine fishes, under 30 centimetres (12 inches) in length, slender, streamlined, and rather nonspecialized in body form; a few species exceed 50 centimetres in length. The wolf herring, *Chirocentrus dorab,* is exceptional in size among the clupeiforms; this species reaches 3.6 metres (12 feet).

Authorities disagree on many aspects of the classification of the order Clupeiformes, which is usually described as including more families than are treated in this article. In a sweeping revision of the bony fishes, the ichthyologists P.H. Greenwood of Great Britain and Donn E. Rosen, Stanley H. Weitzman, and George S. Myers of the United States have restricted the order to the families Clupeidae (herrings, sardines, and allies), Engraulidae (anchovies), Chirocentridae (wolf herrings), and Denticipitidae, a single, little known African species. The last two families are of purely scientific interest; the dominant members of the order, in abundance and therefore in economic importance, are the herrings, sardines, pilchards, menhadens, sprats, anchovies, and anchovetas. Other fish groups formerly included in the Clupeiformes are the tarpons and bonefishes; salmons, trouts, and pikes; and bony tongues and mormyrs.

Distribution

Most clupeiforms inhabit more or less offshore open waters in abundant schools. Although usually considered pelagic (inhabiting the open ocean), in relation to distribution and life history, they are closer to the neritic (coastal) fauna because they do not usually occur in the really open parts of the oceans; rather, they stay close to shore and in bays. Even the truly pelagic and migratory species spawn close to shore. The geographical distribution of the order is limited mainly by temperature and salinity. About 70 percent of the species occurs in tropical waters, only few visiting subtropical regions. More than 20 species are limited to purely boreal and subarctic distribution. Remarkably few species are found in the Southern Hemisphere. In relation to salinity clupeiform fishes represent a fairly mixed group: most of them, approximately half of the living species, are wholly marine; a smaller part are anadromous (living in the sea but entering freshwater to breed); and nearly the same number are wholly freshwater fishes. The order includes some marine genera with large numbers of species, such as *Sardinella* and *Harengula,* which together comprise more than 60 genera and nearly 220 species. There are fewer anadromous clupeids, about 10 genera with 40 species, distributed mostly in temperate regions, but some in subtropical areas. Freshwater clupeiform fishes include 31 species in 16 genera, most of them limited to the tropics. Nine genera with 15 species inhabit the rivers and lakes of Central and West Africa; six species are distributed in freshwaters of the Indo-Malayan Archipelago and Australia; two genera with four species occur in freshwaters of India; some species of the genera *Sigualosa* and *Dorosoma* occur in Central America; and single species of otherwise marine genera are found in the Amazon River (*Rhinosardinia amazonica*), in the rivers of Borneo (*Ilisha macrogaster*), and in freshwater lakes of the Philippines (*Harengula tawilis*). A few other species occasionally occur in freshwater.

Of the families and subfamilies of the Clupeiformes, the subfamilies Dussumieriinae (round herrings), Clupeinae (typical herrings), and Pristigasterinae, and the family Chirocentridae are purely marine; the Denticipitidae and Pellonulinae are limited to freshwater; the Alosinae (shads and alewives) and Dorosomatinae (gizzard shads) are anadromous, freshwater, brackish, or marine; and the Engraulidae are brackish or marine.

NATURAL HISTORY

It is virtually impossible to make a general statement about the biology of clupeiform fishes, except to say that it varies greatly from one species to another. The life history of the majority of species remains little known. Species of economic importance have been extensively studied in order to discover the biological peculiarities that have the determining roles in abundance and distribution; knowledge of such characteristics, of course, is necessary for efficient fishing.

Spawning habitats

Reproduction. Most clupeiforms lay their eggs near shore, often close inshore or in fjords and bays. Few clupeiform species spawn far from shore or in the open sea, except, notably, the Atlantic herring, which spawns on offshore banks. The majority of the spawning grounds are limited to shallow waters ranging from slightly below mean low-tide level to a depth of about four metres (about 13 feet). Some forms, such as the Atlantic herring, however, do spawn at depths of 40 to 200 metres (approximately 130 to 660 feet). The bottom of the spawning grounds, especially those of species with sticky eggs, tends to be clean, hard, and covered with gravel and sand. Spawning takes place above a soft, muddy bottom only if there is a vegetative cover. The freshwater and anadromous clupeiform species spawn in currents of riverbeds with a low mineral content, in shallows of big lakes, and (less often) in river arms and riverine lakes.

The majority of clupeiform fishes have pelagic (free-floating) eggs, which float in the surface or bottom water layers. Egg position is maintained by the presence of a large swollen space between the egg itself and the outer membrane. Some forms (*Clupea, Pomolobus*) have sticky eggs with an adhesive secretion, so that they stick to stones, gravel, or plants shortly after being released. Freshwater and anadromous clupeiforms usually have eggs slightly heavier than water. Such eggs, which would normally sink to the bottom, are constantly lifted by the slightest current and turbulence resulting from wave action and convection of the water. In rivers they freely drift downstream above the bottom. Only a few freshwater forms, such as the freshwater sardine (*Clupeonella abrau*), have eggs that develop in the surface water.

The number of eggs produced varies greatly, but, in general, smaller species produce few eggs, larger species produce many. One of the smaller sprats (*Sprattus sprattus phalericus*), with a maximum size of eight centimetres (about three inches), produces about 2,000 eggs; one of the biggest shads, *Alosa kessleri kessleri*, can produce more than 300,000 eggs; and menhaden (several species of *Brevoortia*) produce more than 500,000. Freshwater species usually have more eggs than marine species of comparable size, evidently an adaptation against the higher mortality in riverine conditions. Those species of clupeiforms with adhesive eggs produce more eggs than do those with free-floating eggs. Apparently, eggs that develop sticking to the bottom have a much higher mortality rate from predators than do eggs that develop floating in the surface water. Of great importance in reducing mortality rates is "repeated portion" spawning: in the majority, if not in all, clupeiform fishes, the eggs in the gonads do not become ripe all at once but in two or more portions, allowing more eggs in the limited space of the body cavity and enhancing the chances of some surviving if the first are destroyed. The many causes of spawn mortality range from those of a physical character, such as wave action and sudden temperature drops, to biological ones, such as predation by gulls and ducks. An important protective mechanism against destruction of the abundant schools is the remarkably early age at which they first breed; females begin frequently to spawn only a few months after hatching. This, coupled with high fecundity, gives the order a high reproductive potential.

Variation in hatching times

The duration of egg development varies from a few hours to nearly two months. An important factor in the rate of development is the temperature of the surrounding water; the cold-water herrings have the longest developmental period. The egg development of the Atlantic herring takes as long as 47 to 50 days at a temperature of 0.1° C (just above 32° F) but only eight days at 19° C (66° F). Some shad eggs develop in about 75 hours at 17° C (63° F) but require only 49 hours at 19° C. The eggs of the Tanganyika sardine (*Stolothrissa tanganicae*), an open-water, freshwater, surface spawner, hatch in 24 to 36 hours while constantly sinking from the surface to a depth of 75 to 150 metres at 25° C (77° F).

Growth and mortality. The thin, threadlike, newly hatched larva has a shape characteristic of nearly all clupeiform fishes, but its behaviour varies greatly, depending on the habitat. In marine species such as the Pacific herring (*Clupea pallasii*), the larvae, shortly after hatching, tend to be concentrated near the surface and usually stay a long time in the area of the spawning ground. The larvae of the Atlantic herring at first tend to make short, upward movements from the spawning beds on the bottom, then sink back again. They start to make horizontal movements within two hours after hatching and after six hours start to form swarms. As their length increases, the vertical movements become more and more pronounced, particularly at night. Larvae have been found to be dispersed by currents at depths of from one to 600 metres (roughly three to 2,000 feet). Later, juveniles drift with the current on the surface, sometimes as far as 1,300 kilometres (slightly more than 800 miles). The larvae of the Tanganyika sardine, less than two millimetres (0.08 inch)

From (adult Atlantic menhaden) A.H. Leim and W.B. Scott, *Fishes of the Atlantic Coast of Canada* (1966), Fisheries Research Board of Canada, reproduced by permission of Information Canada, and (adult *Anchoa mitchilli*) A.J. Mansueti and J.D. Hardy *Development of Fishes of the Chesapeake Bay Region An Atlas of Egg, Larval, and Juvenile Stages*, by the Natural Resources Institute, University of Maryland

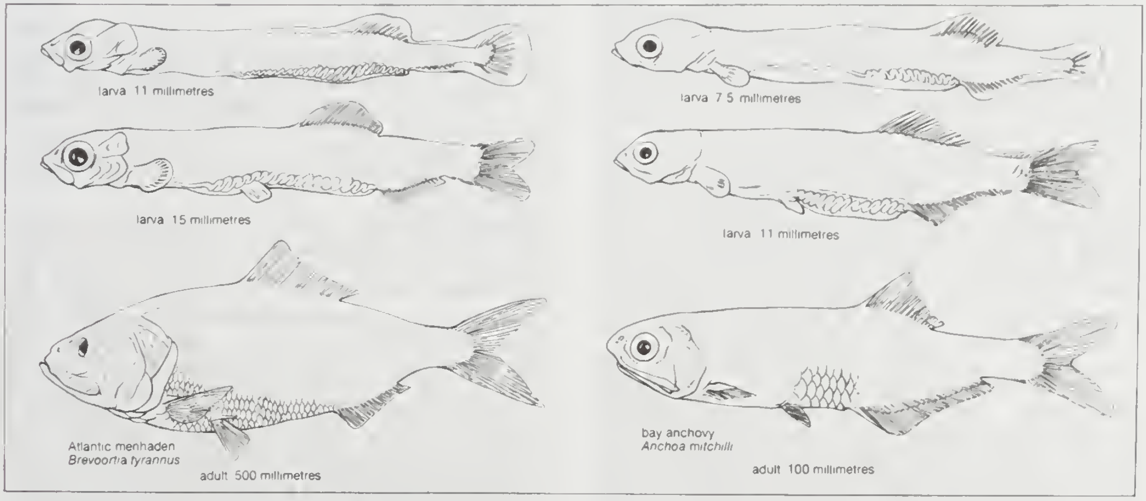


Figure 20: Changes in body proportions and fin position between larvae and adult in two clupeiform fishes.

long, tend after hatching to come straight up by swimming movements of the tail, which is the only flexible part of the body; they sink, however, as soon as they stop movement. Such vertical movement is vital, because the larvae would not survive were they to sink below the level of oxygenated water (80–200 metres [roughly 260–650 feet]). As they grow, the larvae gradually move to the surface waters; when they are about five to six millimetres long, they move toward the shore. They form schools when about 10 millimetres (0.4 inch) in size. In the riverine-spawning Atlantic menhaden (*Brevoortia tyrannus*) the newly hatched pelagic larvae first drift downriver between fresh and brackish water and shoreward from spawning areas and into estuarine nursery areas. Later, pelagic juveniles tend to move upstream as far as 50 kilometres (31 miles), emigrating into the sea only after nearly one year.

Causes of larval mortality

In the early stages of life, all clupeiforms are subject to a high mortality rate, by predation of larger fishes, birds, comb jellies (ctenophores), and arrowworms (chaetognathans) and by being carried out of sheltered bays into localities in which the proper food is lacking. Mortality has been estimated at well over 99 percent, but, because of the extremely high fecundity, the distribution, and the early maturity, the recruitment of new breeding individuals remains high. The age of first sexual maturity is seldom more than three years, and the length at maturity rarely exceeds more than 15 centimetres (approximately six inches). Late-spawning species are usually larger and move over long distances. The age at first breeding is broadly correlated with rate of growth of the individual and with maximum length attained by the species, but there are other determining factors, some of which are unknown. The Siberian shad (*Alosa saposhnikovi*), Baltic sprat (*Sprattus sprattus balticus*), and the *Clupeonella engrauliformis* all mature at two to three years of age but at lengths of 160 to 200 millimetres (roughly 6¼ to 8 inches), 120 to 130 millimetres (4¾ to 5 inches), and 85 to 100 millimetres (3 5/16 to 4 inches), respectively. Different populations of a species may vary in their growth rates; the races of the Atlantic herring vary from two to seven years in age at maturity and from 100 to 185 millimetres (4 to 7¼ inches) in length at maturity. Members of anadromous populations of the alewife (*Alosa pseudoharengus*) reach maturity at three to four years of age and 150 to 170 millimetres (roughly 6 to 6¾ inches) in length, but those of landlocked populations breed at one to two years of age and 95 to 100 millimetres (3¾ to 4 inches).

Migration. During their life cycle some clupeiforms undertake very long migrations of several thousand kilometres; others live in a more or less circumscribed area. Such differences occur, however, even within a species; some races of the herring, for example, spend their entire lives in more or less limited areas; others undertake some of the longest known migrations. Some forms of the Caspian shad (*Alosa caspia*) remain all year round in the southern region of the Caspian Sea, but others move long distances from winter habitats in southern parts to spawning grounds in the northern region of the Caspian.

In addition to spawning migrations, some species travel long distances for feeding. Japanese pilchards (*Sardinella sagax melanosticta*) winter and spawn in the southern part of the Sea of Japan and on the Pacific side of the southern islands of Japan, then move in early summer to the northern end of the Tatar Strait and, in warm years, even to the eastern shore of the Kamchatka Peninsula. Similar or even longer migrations are made by the Californian pilchard or Spanish sardine (*Sardinella anchovia*) and others. Most of these spawning and feeding migrations are from south to north and occur along the coast with the aid of some of the larger ocean currents. As the fish move fairly close to shore, they become the object of intensive fishing.

Long-distance movements of herring

Some of the longest migrations extend over several years and start in the larval stages. The majority of the young Pacific herring spend part or the whole of their first year in shallow coastal waters. Larvae of the Murman race of Pacific herring and Norwegian race (or spring race) of North Atlantic herring usually hatch on offshore spawning grounds and start their long journey drifting with the currents. Those of the Murman race drift with the North Atlantic Current along the coast of northern Norway, north and east, and later, as juveniles, spread actively into the Barents Sea and even into the White Sea. After their first spawning, the Murman herrings move north to the waters around Spitsbergen. The movements of the Norwegian spring herring are similar to those of the Murman race. The young herrings move into deeper water and, as they grow bigger, move farther and farther from the coast. While still immature, they are taken by fisheries in Norway, Denmark, and Scotland and are processed for oil and into meal. As a rule, migrations are oriented by the sea currents near the spawning grounds, but the fish go as well with or against the current direction; four forms of the Caspian shad are known to move against currents.

Food ecology. Intensity of movement and feeding habits determine the relative abundance of various species of clupeiform fishes; these same factors determine economic importance. All of the abundant (and economically important) species feed on plankton—pelagic protozoans (diatoms and flagellates), copepods, metazoan larvae, euphausids, and amphipods. Some apparently feed all year round, as long as food is available, but most change their feeding habits seasonally. It is known that all forms of the herring and most members of the genera *Alosa* and *Clupeonella* do not feed during the spawning season; feeding is most intensive in the summer after spawning and less so in spring before spawning.

Factors affecting abundance

Predatory clupeids seem to be relatively scarce and usually have a much smaller commercial value than do the plankton feeders. The fish-eating race of Russian shad *Alosa kessleri kessleri,* for instance, is far less abundant and is caught less often than is the plankton-feeding race *A. k. volgensis.*

Some evidence suggests that even among plankton-feeding clupeiform fishes, while some species are as a rule abundant, many others are more or less rare. This variation is apparently primarily determined by the size of the inhabited area and the size of spawning grounds; of secondary importance are the time and distance of migrations preceding the age of first reproduction. The Pacific sardine (*Sardinops sagax*)—which inhabits vast areas on both sides of the North Pacific, the South Pacific coasts of South America and Australia, and the Indian Ocean coasts of Australia and Africa—is a good example of a widespread, highly migratory, and economically important species; the herring *Clupea harengus* provides a similar example. Most of the Pacific races of herring, on the other hand, are local and nonmigratory, and their role in commercial catches is far below the value of the Atlantic races. The Japanese pilchard is known to feed in southern as well as in northern regions, and from the ecological point of view this whole area of the Pacific is fully utilized. The high abundance of anchovies is determined more by their early age of sexual maturity than by their movements; similarly, the relatively high abundance in a restricted habitat of the Tanganyika sardines appears to stem from precocious breeding.

The size of the inhabited area is reflected in the presence of more progressive adaptive morphological characteristics. All of the clupeiforms with more primitive features (*e.g., Denticeps, Dorosoma, Clupeonella*) are less abundant and are limited to small areas. Tropical genera have more different species; subtropical and temperate genera are more often monotypic (comprising a single species) but far more abundant.

Schooling behaviour. With few exceptions, the important behavioral characteristics of clupeiforms are schooling and diurnal (daily) vertical movements. Schools are formed with larvae or young juveniles. A fish less than 10 millimetres long approaches the tail of another; both vibrate their bodies in a series of rapid motions, after which they swim together. Occasionally, they are joined by others, and, as the fish grow a few more millimetres in length, the first small schools increase in size and begin to show a steady schooling pattern. Opinions differ on whether the school keeps together through visual contact—it sometimes tends to break up at night—or through sensations received by the lateral-line system, a series of sensory endings extending along the side of the fish. When the schools do persist after nightfall, the lateral-line system may also play a significant role in preventing one animal from straying.

Size, shape, and arrangement of schools

Single schools of herring or anchoveta have been estimated to include many millions of individuals, and some authorities assert that as many as 3,000,000,000 fish may occur in a single school. Even a big school such as this behaves as if it were one organism, with a roughly spherical shape, which is flattened when the school comes into shallow waters or approaches the surface. Within a school of anchovies, the larger individuals tend to be below, the smaller ones above, so that light is allowed to filter through the whole school. There are limits to the size of individuals in any big school; for herring, the difference between the largest and smallest members of a school is about 50 percent. Fishes above or below the size limit break away and form schools among themselves, but even large uniform schools occasionally break apart, and small schools may fuse into larger units. The uniform size of individuals within a school (mostly the same age group) is of great convenience to man, as the fish sort themselves out naturally for canning.

Within the school each fish usually is spaced evenly with enough room between it and the others to swim but not to turn around. In all schools of some species, and in some schools of others, the fish swim with their heads side by side; in other species (*e.g.,* herring) the head of each fish lies next to the middle of its neighbour's body. The schools may spread out or become very tight, depending on the occasion.

The adaptive value of schooling

The primary advantage of the schooling habit seems to lie in the safety of the individual fish. Sardines react to attacks by predators by swimming closer together and milling around in tight, compact balls; herring form a close school with any approach of danger. The reaction of anchovies to predators is even more intense; a school that may be spread over several hundred metres contracts at the approach of a predator to a moving, writhing sphere of thousands of fishes only a few metres across. In such a situation the predator cannot concentrate on a single individual and may be frustrated in its attempt to catch any fish. The adaptive value of schooling behaviour is poorly understood, but several logical explanations have been advanced. Schooling evidently provides a better chance for small fish to survive many environmental hazards than if they live solitarily. The instinctive tendency of the tiny larva to associate, even though hatched from scattered eggs, ensures the formation of the school, with its protection from predation. Certain hydrodynamic interactions between members of the school are thought to facilitate feeding movements; and the aggregation of so many fish simplifies the finding of mates.

Although anchovy schools progress steadily through the water, they do not seem to have any leader or leader groups. Observers from the air have noted that

> fish travelling in the vanguard often drop back and are replaced by others from the flanks and this is repeated in due course. When the school changes course, the fish from the flank find themselves on the leading edge and the previous leading edge becomes a flank. These manoeuvres are carried out with such precision that one has the impression of watching a single creature moving through the water.

The behaviour of the school is determined most probably by the order of feeding. If a school were to swim straight forward, the fish in front would capture most of the food organisms, and those in the rear would starve. Instead, the leading individuals turn back to either flank and, step by step, return to the rear of the school; in this way, each fish gets its turn to feed.

The depth at which the schools swim depends on the movements of plankton, light intensity, temperature, and the maturation cycle of gonads (*i.e.,* whether or not the fish are in breeding condition). There are diurnal vertical movements of schools, related mainly to the corresponding movements of plankton. Most clupeiform schools are believed to stay near the bottom or in deep water during the day and to move toward the surface during the night. Herring often make a vertical migration from a depth of 300 to 400 metres (roughly 1,000 to 1,300 feet) in the day toward the surface water at night; they therefore move from deep cold waters of about 3° C (37° F) to somewhat warmer surface waters of 5° to 7° C (41° to 45° F). On moonless nights clupeid schools can be concentrated on the surface by beams of strong light, a behavioral pattern often exploited by fishermen.

FORM AND FUNCTION

Distinctive differences in fin position

Distinguishing characteristics. The main differences evident among the various clupeiform groups lie in the positions and sizes of the various fins. If a herring (*Clupetta*), pilchard (*Sardinops*), and sprat (*Sprattus*) are held, each by the leading edge of its dorsal fin, the herring hangs approximately horizontally, because the fin is at the centre of the back; the pilchard hangs with its tail lower, the fin being nearer the head; and the sprat drops its nose, because the fin is nearer the tail. The differences of fin position are not pronounced in the larvae, which have a characteristically elongated form with the dorsal, pelvic, and anal fins located far back. The forward part of the body forms an extremely elongated wormlike feature, and, most characteristic, the dorsal fin is never above the pelvic fins as it is in adults, but is well back, usually somewhere between the pelvic and anal fins; in larval anchovies it is even above the anal fin. During the larval transformation the elongated anterior part of the body becomes progressively shorter, as the fins shift forward by a complicated morphological process. The dorsal fin is shifted forward above the lateral body muscles (myomeres); the pelvic fins move backward to their adult position; and the anal fin moves forward simultaneously. In adults of the families Denticipitidae and Chirocentridae, the dorsal fin stays above the anal fin, far back on the body; in the Engraulidae it usually stops a little farther back than the pelvic fins; and in the Clupeidae it generally reaches a position directly above the pelvic fins. As a rule, however, even within families and genera the relative positions of the dorsal, anal, and pelvic fins are somewhat variable and are often used in classification. The position of the dorsal fin becomes stable at the time the larvae transform into juveniles. The positions of the anal and pelvic fins, however, often change later in life, probably because of the swelling of the body cavity with gonad development.

With only a few exceptions, fishes with more forwardly positioned dorsal fins have fewer rays in their anal fin but more rays in the dorsal. The lateral line canals on the head are most developed in fishes with the dorsal fin located anteriorly. The lateral line system serves as an orientation device. As it is sensitive to disturbances in the surrounding water, it is most important in fishes that school densely. Not surprisingly, the species with the most progressively developed morphological features (*i.e.,* the greatest changes from the "primitive" condition of the larval stage), such as the anteriorly located dorsal fin, a smaller number of rays in the anal fin, and a strong lateral-line system on the head, are the best swimmers and undertake the longest migrations.

Relative numbers of denticles and teeth

The development of denticles (toothlike skin projections) and teeth represents another specialization of evolutionary importance. The most primitive clupeiform fishes have an enormous number of dermal denticles (on the head and in the mouth), which have been replaced in evolutionally more advanced forms by teeth, which are larger and fewer in number. In *Denticeps,* for example, the whole head and part of the body are covered by numerous small dermal denticles. Different species of the Clupeidae have small denticles or teeth limited to the bones of the mouth cavity, and anchovies have rows of tiny teeth in the jaws. Finally, *Chirocentrus* has straight sharp teeth on the upper jaw, the tongue, and in a few other places in the mouth, and has large "canine" teeth on the lower jaw.

The value of the ventral keel

The ventral part of the body in the majority of clupeiform fishes forms a keel, the function of which is widely considered to be an adaptation for removing the sharp shadow that would be created below the central part of the body by top lighting, were the fish cylindrical. Prevention of such a shadow is important to an open-water fish often living close to the surface and unprotected from all sides. Seen from below, the keel and the glossy silver

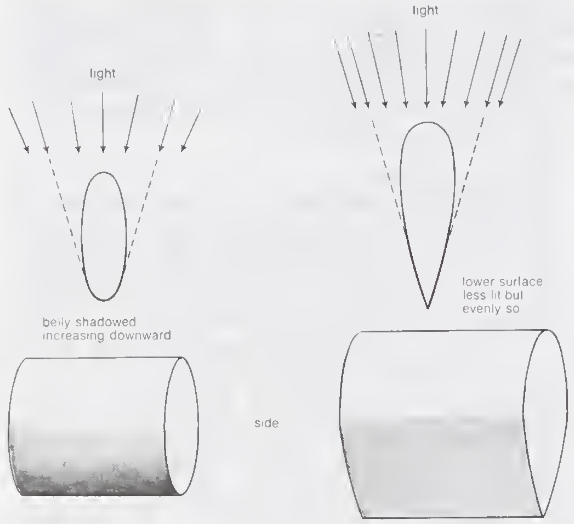


Figure 21: Value of the keeled belly in surface-swimming fishes (see text).

sides of the body cause the fish to disappear in the mirror-like reflection of the water surface. Viewed from above, the fish is protected by the dark cryptic colouring of the dorsal part, which simulates the colour of the deep water. The predator who encounters and sees the whole school is also deceived by the resemblance of the tight school to a larger organism. Against man's nets and electronic devices, however, such coloration and schooling behaviour afford little protection.

Physiology. The movement of anadromous clupeiforms from the salt ocean into freshwater rivers and lakes requires special physiological adaptations to regulate the blood's osmotic pressure (basically, the pressure of a water solution of salts exerted in either direction against a semipermeable membrane caused by differences between the concentrations of dissolved salts within the body and those outside, in the sea). When a fish enters water of salinity lower than seawater, slight increases in osmotic pressure cause the kidneys to excrete larger amounts of water. The conversion from saltwater to freshwater physiology requires some time, however, so the fish usually remains in brackish waters to avoid a sudden physiological shock. During the periods when anadromous fishes are migrating into or out of freshwater, they form large aggregations in estuaries, awaiting the changeover in their osmotic regulating systems.

CLASSIFICATION

Distinguishing taxonomic features. Three main character complexes have recently been recognized and accepted as distinguishing the clupeiform fishes: (1) the presence of an internal connection between the swim bladder and the inner ear, usually forming two large vesicles (cavities) within the skull bones; (2) certain peculiarities of the skull, involving the relation of the lateral line canals to each other and to the ear; (3) certain complex features in the caudal (tail) fin skeleton.

Annotated classification. A recent and widely accepted classification of the order Clupeiformes by P.H. Greenwood *et al.* is presented below. These authors consider the Clupeiformes as the only order in the superorder Clupeomorpha, which formerly included other major fish groups (see above).

ORDER CLUPEIFORMES

Silvery, laterally compressed fishes; mainly marine, but many anadromous or wholly freshwater; mostly pelagic and schooling fishes. Lateral-line canal on head usually extending over operculum (gill cover). About 400 living species.

Suborder Denticipitoidei

Caudal skeleton of extremely primitive type; small arches present on 2 centra (bodies of vertebrae) to carry the first 3 hypural bones (fused spines of the vertebrae) of the tail fin. One family.

Family Denticipitidae

The most primitive living clupeiform. Numerous dermal denticles present on head, on the dorsal part of the secondary pectoral girdle, and on the scales around the anterior end of the lateral line. Lateral line completely developed on the trunk. A single living species, *Denticeps clupeoides*, in fast-running clear water in medium-sized streams of West African Nigeria; and a single fossil species, *Palaeodenticeps tanganikae*, from the Tertiary lacustrine sediments in East African Tanzania.

Suborder Clupeoidei

Characteristic caudal skeleton: the second hypural bone lacks any connection with the urostyle (tail support) and is separated from it by a distinct gap. Lateral line pores completely lacking on trunk. Keeled scutes (projecting scales) usually present along the ventral midline of the abdomen.

Family Clupeidae (herrings, sardines, pilchards, shads, menhadens, and allies)

Teeth usually absent in mouth or very weakly developed; minute in jaw. Keel scales well developed, except in round herrings (subfamily Dussumieriinae), in which they are absent, and the ventral part of body rounded. About 50 genera and 190 species, virtually worldwide in marine waters and in many bodies of freshwater.

Family Engraulidae (anchovies)

Mostly smaller fishes than clupeids, with the snout projecting beyond the very wide mouth. Upper and lower jaws usually armed with rows of minute teeth that sometimes become larger in the posterior end of the jaws. About 200 species; primarily marine with a few anadromous; found in very large schools.

Family Chirocentridae (wolf herrings)

Body laterally compressed and elongated, with sharp, keeled ventral margin; scales small. Lower jaw strongly projecting; large fanglike teeth in both jaws. Two species, *Chirocentrus dorab* and *C. nudus;* widely but sparsely distributed in the Sea of Japan, the Pacific Ocean off Australia and in Melanesia, the Red Sea, and along the east coast of Africa. Used for food in some areas but not very palatable. Larger than other clupeiforms, reaching at least 3.6 metres (12 feet) in length.

Critical appraisal. Until the revision of the bony fishes by Greenwood and his colleagues in 1966, the most widely accepted classifications were those by an American, C.T. Regan, in 1929, a Soviet ichthyologist, L.S. Berg, in 1940, and two from France, L. Bertin and Camille Arambourg, in 1958. The three earlier systems differ widely from one another in the scope of the order Clupeiformes, in the subdivisions of the order, and in the order of families, but they have in common the inclusion of many more groups than were considered related to the clupeid fishes by Greenwood *et al.* The earlier classifications grouped together in one order, Clupeiformes or Isopondyli, a large number of fishes characterized by having soft, as opposed to spiny, fin rays.

Greenwood *et al.* postulated, on the basis of a number of other features in both modern and fossil fishes, that this similarity is overridden by more fundamental differences that indicate a long history of phyletic separation. These authors separated the families Denticipitidae, Clupeidae, Engraulidae, and Chirocentridae in a distinct superorder Clupeomorpha, placed in Division I, one of the three subgroups of the bony fishes. The bony tongues, mormyrs and relatives, treated by Bertin and Arambourg as suborders of the Clupeiformes, were placed by Greenwood *et al.* in the superorder Osteoglossomorpha, sole group in their Division II. The remaining fishes formerly included in the Clupeiformes, consisting mainly of the salmons, trouts, pikes, and a number of deepsea forms, were placed in a large order Salmoniformes, part of Division III. (E.K.B.)

Bony tongues, freshwater butterfly fishes, and allies (Osteoglossomorpha)

The superorder Osteoglossomorpha is a group of morphologically and biologically diverse primitive fishes primarily found in freshwaters; a few species enter slightly brackish water. Their relationship with other teleosts (*i.e.*, advanced bony fishes) is obscure; they probably were an early offshoot from the basal teleost stock. Osteoglossomorpha comprises six extant families and about 150 species. Although the group is of little importance to man, in parts of Africa, Asia, and South America certain osteoglossomorph species are sometimes sought commercially as food fishes.

Except for one North American family (Hiodontidae), the Osteoglossomorpha are tropical fishes. The families Mormyridae (elephant-snout fishes, mormyrs), Gymnarchidae, and Pantodontidae (butterfly fishes) are confined to Africa; the Notopteridae (featherbacks) occur in Africa, Southeast Asia, and India. The distribution of the Osteoglossidae (*e.g.*, pirarucu, arawana) in Africa, South America, and Australasia (believed by many authorities to have once been joined as a single landmass called Gondwana) is of particular zoogeographical interest.

The pirarucu (*Arapaima*) of the Amazon, one of the world's largest freshwater fishes, attains a length of three metres (about 10 feet); other osteoglossomorphs—for example, certain mormyrids—are only a few centimetres long.

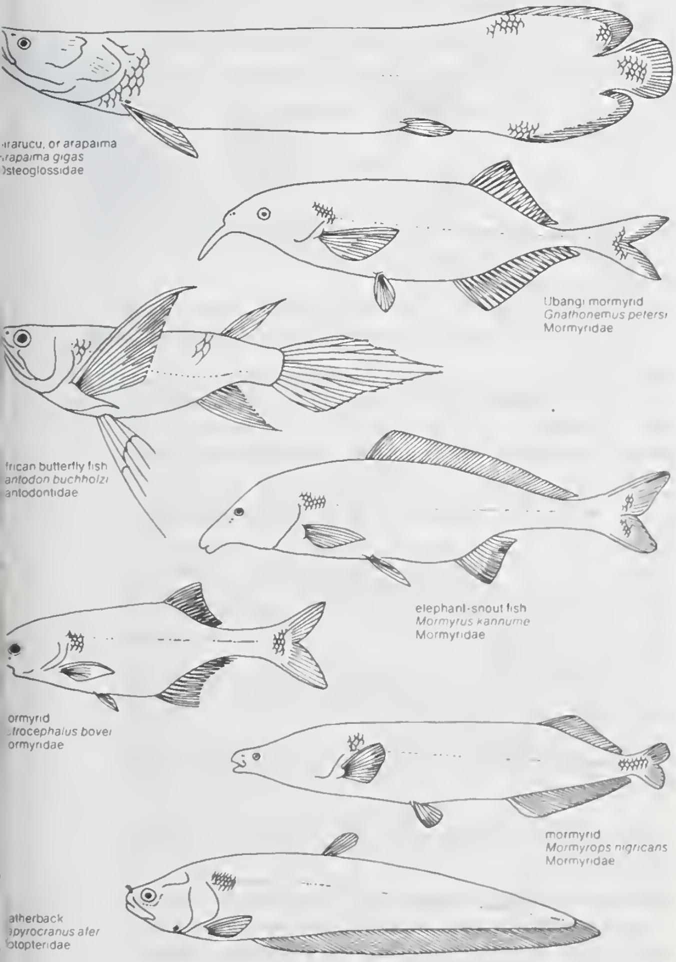


Figure 22: Representative osteoglossomorphic fishes.

NATURAL HISTORY

Life cycle and reproduction. A variety of breeding habits have evolved among the Osteoglossomorpha. In some species there is considerable care of the young by the parents. Although no species is known to undertake extensive breeding migrations, many leave the usual habitat and move into floodplains or streams at breeding time.

The breeding biology of the Mormyridae has been little studied; it does not seem likely, however, that they prepare spawning nests or exercise much parental care. In contrast, *Gymnarchus niloticus* (Gymnarchidae) prepares a large floating nest from the matted stems of swamp grasses, biting off the stems and fashioning them into a trough-shaped structure with an internal length of about 50 centimetres (20 inches). Spawning takes place in the nest, and one or both parents guard the developing young for approximately 18 days.

Nests are not made by notopterioids, but they do establish a breeding territory. Both *Hiodon* species (goldeye and mooneye) spawn in the spring. Eggs are laid on gravel or rocks in shallow, quiet water, which the adults reach after a short migration. The young of the goldeye (*H. alosoides*) remain there until late summer before migrating downstream. In *Notopterus chitala* (Notopteridae), one parent, probably the male, clears an area of the bottom near some submerged object (*e.g.*, rock, plant stem, or piling) on which the eggs are later laid in circular bands; the male guards the developing embryos.

Care of the young

Members of the Osteoglossidae care for their young in a variety of ways. The South American *Osteoglossum bicirrhosum* and its Indo-Australian relatives *Scleropages leichardti* and *S. formosus* carry the eggs and young in the mouth of one parent; little else is known of their breeding habits. The African *Heterotis niloticus* prepares a crude nest from grasses in newly flooded swamp plains. The male guards the young and leads them from the nest on feeding excursions. Both sexes of *Arapaima gigas* of South America dig a spawning pit and guard the developing embryos, which hatch and leave the pit after about seven days; care is provided by the male, around whose head the young congregate. The dark colour of the male's head seems to provide the chief stimulus for shoal orientation; *i.e.*, forming groups or schools. Apparently there is also a gustatory (*i.e.*, taste) or olfactory (*i.e.*, smell) stimulus in a secretion from glands on the male's head. Parent–young groups persist for two to three months; if the male dies, the young join other shoals.

Spawnings in aquaria suggest that *Pantodon* (Pantodontidae), the African butterfly fish, produces floating eggs and provides minimal parental care. A complicated courtship in this species has also been observed.

Electrical organs

Behaviour and ecology. Mormyridae and Gymnarchidae are of particular interest because they have electrical organs. Electrical discharges from these organs are in the form of pulses, the frequency and nature of which are different in each species. In nature, electrical discharges ranging from an output of about 120 to 300 pulses per second have been recorded.

The flow pattern of the electric field around the fish is distorted by the different conductivity of objects that pass into it. These variations are detected by modified nerve cells (mormyromasts) in the skin. The greatly varying conductivity differences among animals and inorganic objects make it possible for mormyrids to use their electrical organs to distinguish between prey, predators, and obstacles in the turbid water they often inhabit. Discharges from the organs also serve as signals to other mormyrids.

Some mormyriforms tend to swim with little body movement, using instead the dorsal fins for propulsion. This unusual swimming method is probably associated with the use of electric organs in navigation and detection; *Gymnarchus*, for example, swims with its body held straight, propulsion being provided by undulations of the dorsal, or back, fin. Since electrical organs lie near the tail, side-to-side movements of the tail end (as in normal swimming movement) would constantly change their position relative to that of the receptor organs, which are in the head area.

Rigid-bodied swimming like that of the mormyriforms also occurs in the featherbacks (Notopteridae), which use the long anal fin for propulsion. There is no electrical organ in the notopterids, however; the rigid body of these fishes may be correlated with the long, gas-filled swim bladder that extends into the tail end of the body.

All other osteoglossomorphs swim by using the body musculature and caudal (tail) fin in the usual manner. The little African butterfly fish (*Pantodon*) has greatly expanded winglike pectoral fins (behind the gills), which are used for short flights in the air, either to escape predators or to catch insects. *Pantodon* habitually swims or drifts just below the water surface. It leaps from the water by means of a powerful thrust of the pectorals, sending the fish 30 centimetres (about a foot) or more vertically out of the water. Short horizontal flights of about one metre (40 inches) are also executed.

Some species of Osteoglossidae (*Heterotis, Arapaima*), the related Pantodontidae, certain Notopteridae (*Notopterus, Papyrocranus, Xenomystus*), and the Gymnarchidae are able to breathe air at the surface; thus they can live in areas where the water is deoxygenated.

The notopterid *Xenomystus* produces sounds that are

The swim bladder, noise-making, and hearing

used as warnings and in courtship. Swim-bladder structure in other Notopteridae suggests that they are also capable of emitting sounds. In all Notopteroidei (including the Hiodontidae) the swim bladder is closely connected with the inner ear, a condition that may be an aid to hearing. Except for the osteoglossid *Heterotis*, all the osteoglossomorphs are carnivorous, the smaller species (and young of all species) feeding on insects and other invertebrates, the larger species on fish. *Heterotis* feeds on microscopic plants and animals filtered from the water.

Osteoglossomorph fishes occupy a diversity of habitats in rivers and lakes, often in turbid waters or in regions with dense aquatic vegetation. A few species seem to require open waters, and some Notopteridae can tolerate slightly brackish water.

FORM AND FUNCTION

All living Osteoglossomorpha have strongly toothed jaws. The jaws are not protrusible, but in piscivorous (*i.e.*, fish-eating) species the gape is sometimes considerable. The Mormyridae show a great variety of head shape and mouth form. Many insectivorous (*i.e.*, insect-eating) species have a small mouth at the tip of a long, curved, tubular snout; they feed by probing among rocks or in soft mud. Piscivores have larger mouths and short snouts.

The mouth is large in all Osteoglossidae, which, like the Notopteridae and Hiodontidae, have a more typical head shape than do the Mormyridae. *Heterotis* has complex spiralled structures lying above the gill arches on each side of the head and opening into the pharynx. Food particles, drawn into the pharynx with the respiratory current, are filtered out and concentrated in these organs.

In all Mormyridae and in the Gymnarchidae, a short length of body musculature toward the tail is modified to form the electric organ. These muscles have lost the ability to contract and have undergone considerable cellular reorganization, from slender fibres of ordinary muscle into thin, flat, electroplates, the structures in which electricity is produced. There are about 300 to 400 electroplates in mormyrid electric organs, and 600 to 800 in those of *Gymnarchus*. The brain, particularly the cerebellum, of mormyriforms is the largest known among fish; the cerebellum is associated with the electroreceptor organs (mormyromasts) in the skin.

A well-developed swim bladder is present in all Osteoglossomorpha. In the air-breathing osteoglossids, pantodontids, and gymnarchids, it is highly vascularized (*i.e.*, has many blood vessels), and its inner surface is honeycombed with tiny pits.

Paired extensions in the Notopteridae and Hiodontidae connect the swim bladder with the auditory region of the skull. In the African genus *Papyrocranus*, diverticula (outpocketings) of the swim bladder actually penetrate the skull, a condition that probably improves hearing. Posteriorly the swim bladder in notopterids extends beyond the abdominal cavity and runs back on either side of the vertebral column. In the early embryos of Mormyridae and Gymnarchidae, a pair of thin tubes extends forward from the swim bladder into the skull; later the tubes atrophy, leaving an isolated vesicle—a blister, or balloon-like structure—within the skull surrounded by the semicircular canals of each ear.

The inner ear shows various modifications; in notopterids, gymnarchids, and mormyrids, the upper portion (for balance) is completely separated from the lower part (for hearing).

In the Mormyridae and Pantodontidae, the anal fin shows considerable sexual dimorphism in shape—*i.e.*, the anal fin of males differs from that of females. These differences may be related to spawning activity.

CLASSIFICATION

Distinguishing taxonomic features. Classification in the superorder Osteoglossomorpha is based largely on skeletal characters, in particular the caudal-fin skeleton; the bones around the eye; and the gill arches and their associated dentition. Details of the inner ear and swim-bladder anatomy are also of importance.

Annotated classification. The classification used here is based on that of P.H. Greenwood, D.E. Rosen, S.H. Weitzman, and G.S. Myers (1966), with subsequent research by Greenwood (1970) and G.J. Nelson (1968, 1969). Groups indicated by a dagger (†) are extinct and known only from fossils.

SUPERORDER OSTEOGLOSSOMORPHA
Primitive; well-developed teeth on tongue, skull base, and bones of the mouth cavity; caudal fin skeleton of characteristic form. Lower Cretaceous to Recent.

Order Osteoglossiformes
Osteoglossomorph fishes without electric organs.

Suborder Osteoglossoidei
Swim bladder not connected with skull; semicircular canals and lower part of inner ear connected.

Family Osteoglossidae. Fishes of diverse body form; pectoral fins not greatly enlarged, pelvic fins abdominal in position. Genera: *Arapaima* (1 species) and *Osteoglossum* (2 species), South America; *Scleropages* (2 species), Australia, New Guinea, Borneo, Sumatra, Malaysia, Thailand; *Heterotis* (1 species), Africa. Fossils from Eocene of North America and Tertiary of Australia (*Phareodus*); Eocene of Sumatra and Tertiary of India (*Musperia*).

†*Family Singididae.* Extinct. Apparently toothless; monotypic genus (*Singida*) from Paleocene of Tanzania (East Africa).

Family Pantodontidae. Greatly expanded winglike pectoral fins; pelvic fins thoracic. A monotypic genus, *Pantodon buchholzi*, from Africa. No fossil record.

Suborder Notopteroidei
Swim-bladder connected with the skull; semicircular canals separate from lower part of ear, or, if connected, utriculus greatly enlarged.

†*Family Lycopteridae.* Extinct. Lower Cretaceous of northeast Asia; small freshwater fishes resembling the Hiodontidae. 4 genera (6 species).

Family Notopteridae. Long anal fin confluent with reduced caudal; dorsal fin small or absent. Genera: *Papyrocranus* (1 species) and *Xenomystus* (1 species) in Africa, *Notopterus* (4 species) in Asia and Indonesia. Fossil *Notopterus* from Eocene of Sumatra.

Family Hiodontidae (goldeye and mooneye). Probably the most primitive living osteoglossomorphs. One genus, *Hiodon*, confined to North America. Fossil from Eocene of British Columbia (*Eohiodon rosei*).

Order Mormyriformes
With electrical organs; very diverse head shape and mouth form. Confined to Africa; fossils from Pliocene of Egypt; 2 families, about 130 species.

Family Mormyridae (mormyrs and elephant-snout fishes). Anal, caudal, and dorsal fins present; several genera; about 130 species.

Family Gymnarchidae. No anal fin; long dorsal confluent with reduced caudal fin. One genus and species, *Gymnarchus niloticus*.

Critical appraisal. Taxonomic problems of the osteoglossomorphs concern intragroup relationships, particularly whether the Mormyriformes are more closely related to the Osteoglossoidei or to the Notopteroidei. Some authorities relate the Mormyriformes with the Notopteroidei; others suggest closer affinity with the osteoglossoids and propose that one order, rather than two, should be recognized. The superorder as a whole presents problems of relationship with the other teleostean lineages. Certain fossil groups—*e.g.*, the Eocene genus *Brychaetus* and the Mesozoic families Plethodontidae and Ichthyodectidae—may be osteoglossomorphs, but their relationship with the living groups is still obscure. *Brychaetus* can probably be classified in the Osteoglossiformes, but it may represent a distinct suborder. (P.H.G.)

Salmons, trouts, whitefishes, and allies (Salmoniformes)

The order Salmoniformes is a diverse and complex group of fishes. The order consists of about 1,000 species in the fresh waters and in the oceans of the world. Included are the familiar trout, salmon, and pike, as well as most of the bizarre forms of fishes inhabiting the middepths of the oceans.

The framework of the order Salmoniformes in the dis-

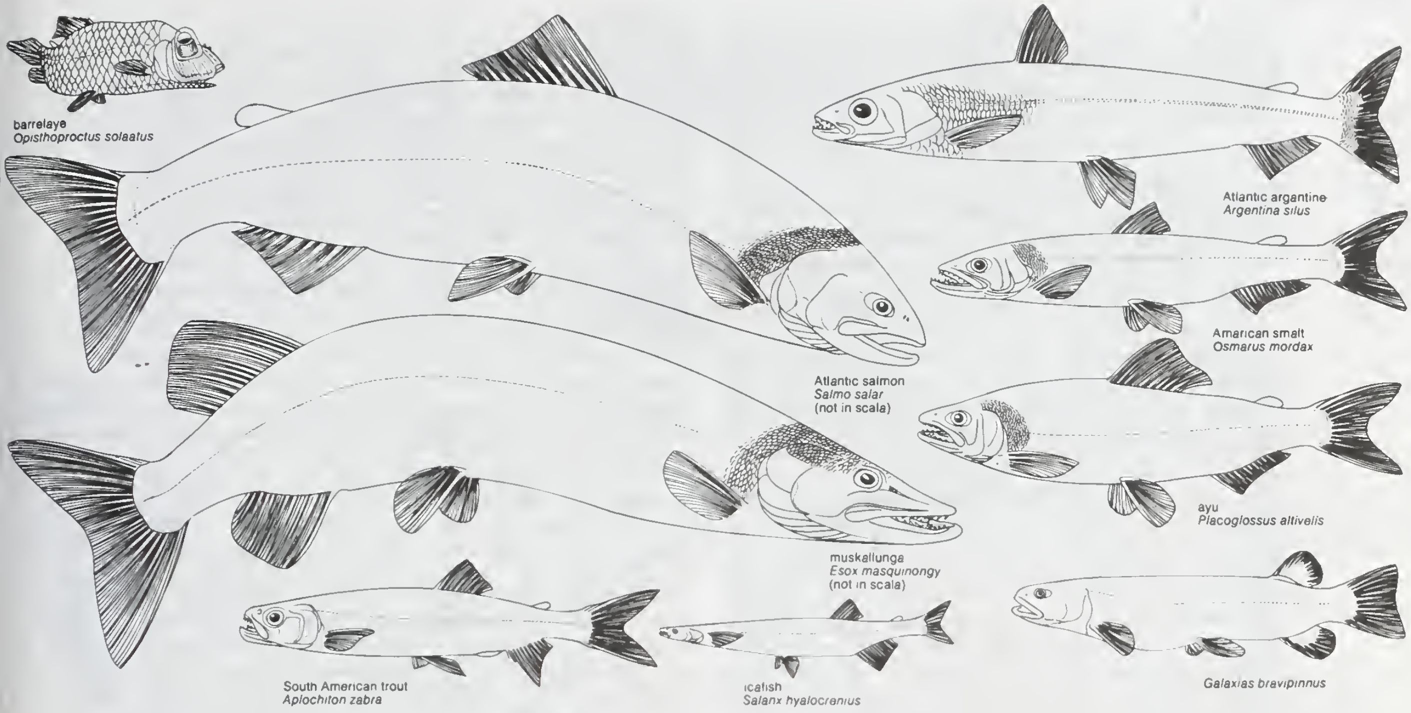


Figure 23: Body plans of representative salmoniform fishes.

cussion below should not be considered as a definitive taxonomic category but rather as an assemblage of diverse fishes possessing several primitive anatomical features representative of an early stage in the evolution of modern bony fishes (teleostean fishes).

GENERAL FEATURES

Evolutionary importance of the order. The significance of the order Salmoniformes as presently classified—a major departure from early schemes of fish classification—is in the evolutionary position of the group; the Salmoniformes are now considered a basal stock in the mainstream of modern bony-fish evolution. The present classification implies that the ancestors of salmoniform fishes developed several evolutionary trends in the Late Mesozoic Era, about 100,000,000 years ago, providing the necessary source of evolutionary raw material to initiate several successful evolutionary lineages, ultimately leading to most of the modern bony fishes.

Divisions of the order

The order Salmoniformes, as treated in this article, is based on the combined work of a British ichthyologist and several American ichthyologists, who have divided the order into eight suborders and 37 families. Constructed in this way, the order Salmoniformes brings together a variety of extremely diverse groups. Several major alterations have recently been proposed for the classification of salmoniform fishes; thus the classification used in this article is provisional and may be subject to extensive changes when more precise information is available. The order Salmoniformes can be considered more as an evolutionary grade in the phylogeny of teleostean fishes than a well-defined taxonomic category. No single character or group of characters can distinguish all salmoniform fishes from all other fishes.

Reasons for interest in the order. The trouts, salmons, chars, whitefishes, and graylings of the family Salmonidae are the most widely known and intensively studied family of fishes. Their famed sporting qualities and excellent taste ensure their economic importance. At the other extreme, some deep-sea families of salmoniform fishes are known only to a few ichthyologists, and often only on the basis of a few imperfectly preserved specimens. The bulk of salmoniform species are fishes of the middepths (mesopelagic and bathypelagic zones) of the open oceans. The deep-sea salmoniforms have evolved unusual body forms and structures—such as luminous organs, telescopic eyes, complex appendages, and enormous and well-developed jaws and teeth—to cope with existence in the twilight and dark zones of the ocean and are of great interest in the study of evolutionary and developmental biology. Some of the anatomical structures evolved by the deep-sea salmoniforms are among the most striking and strange ones found in the animal kingdom.

Size range. The largest of the salmoniform fishes are members of the family Salmonidae and include the Pacific king salmon (*Onchorhynchus tshawytscha*) and the Danube and Siberian huchen (*Hucho hucho*), both of which are known to attain a weight of 50 kilograms (110 pounds) or more. The North American muskellunge (*Esox masquinongy*), a member of the pike family, also approaches this size. The majority of the salmoniform species, however, are small. Most of the deep-sea species do not exceed 150 millimetres (six inches) in length, and many at maturity are no more than 25 to 50 millimetres (one to two inches) long. The largest of the marine salmoniforms is one of the lancet fishes (family Alepisauridae); this species may reach a maximum length of 2.1 metres (about 7 feet). The lancet fish has an elongated body with tremendously enlarged, dagger-like teeth, a fragile sail-like dorsal fin, and a soft, flaccid body. Occasionally a lancet fish migrates from the ocean depths to the surface and may be caught by a fisherman or found washed up on a beach. Lancet fishes, like other deep-sea salmoniforms, are so highly modified that the relationships to trout and salmon are not obvious. A basic salmoniform feature, however—the small, fleshy adipose (fatty) fin that is situated between the dorsal fin and the tail—is found on both trouts and lancet fishes and has been inherited from a very ancient, but common, ancestor.

Most salmoniform species, including the smaller forms, are predacious fishes. Many peculiar modifications have been evolved by the small species of marine salmoniforms to allow them to capture and consume prey sometimes as large as themselves. Some of these lilliputian monsters appear to be mostly head and jaws, all out of proportion to the soft, gelatin-like body. The lack of a neck in fishes limits the mobility of the head and jaws; remarkable adaptations increase the flexibility of the head and allow a great enlargement of the mouth opening to engulf prey quickly. Several deep-sea salmoniforms lack bone on the anterior portion of the vertebral column, increasing the flexibility of the head and jaws.

Distribution and abundance. Salmoniform fishes are found in fresh water on all continents and in all of the oceans of the world. Various representatives of the trout, pike, and smelt families are indigenous to the cooler fresh waters of the Northern Hemisphere. Species of the family Salmonidae inhabit the colder waters of North America, from tributaries of the Arctic Ocean to tributaries of the Gulf of California in northwestern Mexico; in Europe and Asia, a comparable distribution is found, from the Arctic Ocean to the Atlas Mountains in North Africa and to the island of Taiwan. One member of the family, the Arctic char (*Salvelinus alpinus*), is the most northerly occurring of any freshwater fish. The development of an anadromous life cycle—spawning in fresh water but migrating to the sea for feeding and maturation—has allowed species of trout and salmon to extend their range greatly, particularly into fresh waters of glaciated regions after the glaciers recede and waters become inhabitable. The use of marine invasion routes allows a rapid expansion in the distribution of a species into new areas, often inaccessible to other species completely restricted to a freshwater life cycle. Species of the family Salmonidae are clearly the dominant fishes of the recently glaciated fresh waters of the Northern Hemisphere.

The anadromous life cycle

The pike and its allies (family Esocidae) have a distribution somewhat similar to the Salmonidae; however, their range extends neither so far north nor so far south. The pikes are completely restricted to fresh water throughout their life cycle; however, the distribution of the northern pike (*Esox lucius*) in Europe, Asia, and North America is one of the broadest distributional patterns of any fish species. Such a distribution must have been achieved when direct freshwater connections existed between the present major drainage basins and between Asia and North America. The smelts of the family Osmeridae are small fishes of Europe, Asia, and North America. Some smelts are permanent freshwater inhabitants, but the distribution of freshwater smelts is associated with relatively recent geological events; most smelts are anadromous or marine. No smelt species has penetrated far enough inland to establish a broad distribution in fresh water comparable to that of the salmonid fishes. The other salmoniform fishes with anadromous and freshwater species in the Northern Hemisphere are members of the Far Eastern families Plecoglossidae and Salangidae. In the Southern Hemisphere, salmoniform fishes that are ecologically similar to the trouts and smelts are encountered in the fresh waters of southern Africa, southern South America, Australia, New Zealand, and Tasmania. These fishes are classified in the families Galaxiidae, Retropinnidae, Aplochitonidae, and Prototroctidae (of the suborder Galaxioidei). The galaxioid fishes are typically small (measuring only 100 to 300 millimetres [four to 12 inches]) marine and freshwater fishes. The family Galaxiidae contains the most species (about 35) and has the broadest distribution—in Africa, South America, Australia, New Zealand, and Tasmania. The smeltlike fishes of the family Retropinnidae comprise about six species native to Australia, New Zealand, and Tasmania. The family Aplochitonidae consists of three species in southern South America and a larva-like (neotenic) species in Tasmania. The Prototroctidae has two, troutlike species in Australia and New Zealand. Various species of Salmonidae, particularly the North American rainbow trout (*Salmo gairdneri*) and the European brown trout (*S. trutta*), have been widely introduced and successfully established in suitable waters in Africa, South America, Australia and New Zealand. When introduced into lakes with abundant food fishes but previously lacking large predator fishes, the introduced trout flourish, growing rapidly to a large size. In certain lakes in Australia and New Zealand, famed for their trophy-sized trout, the trout feed avidly on their distant relatives, species of the Retropinnidae and Galaxiidae.

The remaining 25 or more families of Salmoniformes, with about 800 species, are entirely marine, typically mid-depth and deep-sea fishes. Representatives are found in all oceans and all depths, but in temperate and tropical waters major concentrations occur at depths of 200 to 2,000 metres (about 650 to 6,500 feet). The tremendous abundance attained by some populations of these marine salmoniforms has become evident only with the advent of modern sonar equipment, which has detected aggregations many square kilometres in extent.

IMPORTANCE

The economic significance of the trouts and salmons both as sporting fishes and as commercial products is well-known. Governments invest heavily to maintain and increase the production of trout and salmon; hundreds of millions of trout and salmon are hatched, reared, and stocked each year for sport and commerce. In fact, a large private industry has developed—particularly in Denmark, Japan, and the United States—to supply trout to markets and restaurants. With the problems of increased human population and the demands made on rivers by industry and agriculture, the challenge of perpetuating and increasing the abundance of salmon and trout has become a serious one for fisheries scientists.

The demand for trout as a sport fish far exceeds the supply in heavily populated regions. This situation, particularly in the United States, has resulted in a massive program by state and federal agencies to raise trout to acceptable size and to stock them in heavily fished waters. Such an artificial abundance, however, is a poor substitute for natural trout fishing.

National programs assisting sport fishing

Except for the pikes, the remaining freshwater salmoniforms are too small or too rare to be significant sport fish, but most are considered excellent food fish. The oceanic salmoniforms have little direct importance to man; because of their tremendous abundance, however, they form a vital link in the food chain of the oceans, providing forage for valuable predator species such as the tuna. Many of the deep-sea salmoniforms undertake daily vertical migrations, rising toward the surface layer of the ocean at night for feeding. This vertical migration exposes them to predation by larger fishes and functions in the recycling of energy in the ocean by elevating energy accumulated in the lower depths (in the bodies of the small salmoniforms) and making it available to large predators in the upper zones.

NATURAL HISTORY

Life cycle and reproduction. Virtually every type of life cycle and mode of reproduction known for fishes is exhibited by some salmoniform fishes. These life cycles range from passage of the entire life-span in the confines of a small pond or stream to migrations encompassing thousands of kilometres from a stream to the ocean and back to the stream. Some species have a direct development stage from the egg, hatching as miniature adults, ready to fend for themselves. Most deep-sea marine species have larval stages, drastically different from the adult. Some larvae have eyes attached to long stalks from the head. Most salmoniform species consist of males and females, but several deep-sea groups are hermaphrodites, a single individual having functional testes and ovaries. Evidently, in the darkness of the ocean depths, it is advantageous for an individual to function both as male and female.

The life cycles of salmons and trouts have been intensively studied because of the economic importance of salmonid fishes. Factual information the life cycle and reproduction is used to settle disputes between nations regarding the origin of salmon caught in the open ocean and for the intelligent management of the resource.

The life cycle and reproduction of the deep-sea salmoniforms, however, are little known except for interpretations gained from examination of a few specimens and collection of eggs and larvae. Eggs and larvae of many of the marine species have not yet been found.

Among the salmoniform fishes, only the pike family (Esocidae) and the mudminnow family (Umbridae) are completely restricted to fresh water throughout their life cycle. All other families that have freshwater representatives contain some species that enter the marine environment for growth and maturation, returning to fresh water to spawn. One species of the family Galaxiidae has a catadromous life cycle—spawning takes place in a marine environment, and the young migrate to fresh water to mature. All of the oceanic salmoniforms are completely marine throughout their life cycle.

The catadromous life cycle

The families Salmonidae and Osmeridae demonstrate a transition between freshwater and marine life cycles. All species of salmonids spawn in fresh water, but the Pacific

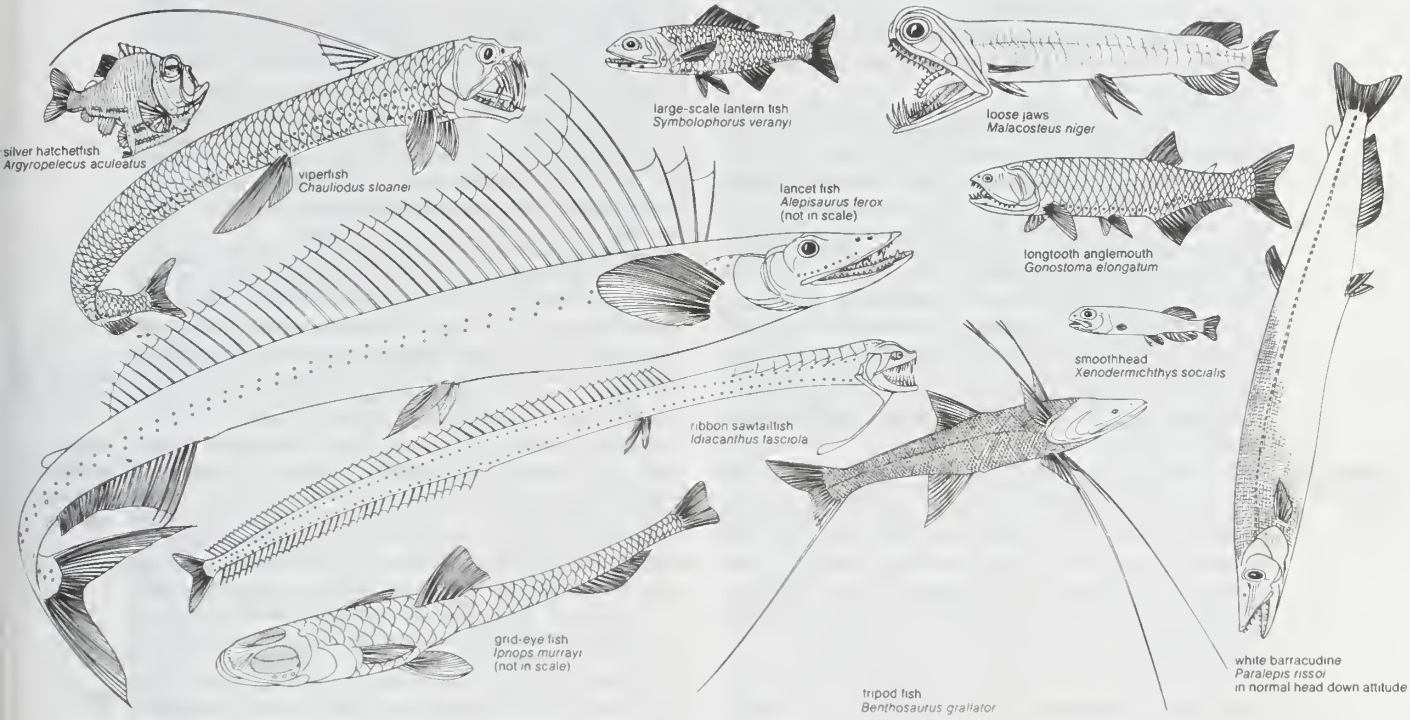


Figure 24: Body plans of representative salmoniform fishes.

pink salmon (*Onchorhynchus gorbuscha*) has reduced the freshwater stage to the spawning migration and incubation of the eggs. As soon as the eggs hatch and the yolk sac is absorbed, the pink salmon fry migrate to sea. Some pink salmon may even spawn in the intertidal zone at the mouths of small streams, virtually eliminating the freshwater stage in the life cycle altogether. Other species of the family Salmonidae, such as the lake char, or lake trout (*Salvelinus namaycush*), the graylings (*Thymallus*), and many of the whitefishes (*Coregonus*), have completely freshwater life cycles. Interestingly, life cycles may differ among closely related species or even between populations of the same species; for example, rainbow trout that go to sea and return as large, silvery individuals are called steelhead trout. A single river system may contain local resident populations of small rainbow trout—maturing, spawning, and completing a life cycle within 100 metres (about 300 feet) of the site of their birth—as well as anadromous steelhead rainbow trout that have returned from the ocean after a two- or three-year journey spanning several thousand kilometres. Evidently the heritable differences that govern the type of life cycle in trouts—anadromous or freshwater—are slight. It has been demonstrated that offspring from anadromous parents can be used to establish populations in completely landlocked environments, and that the progeny of nonanadromous parents may go to sea if given the opportunity.

Reproductive behaviour, the type and size of the eggs laid, and the amount of parental care have been developed in each species by the process of natural selection. In an evolutionary sense, spawning success is ultimately judged by the number of mature adults resulting from any spawning act. If the eggs and larvae are exposed to a harsh and perilous environment, there is a selective advantage for a female to produce fewer but larger eggs and to provide some extra measure of protection for the developing embryos. Cold, swift rivers with sparse food, typically utilized by salmons and trouts for spawning, undoubtedly have been a major selective force in the evolution of large eggs (four to eight millimetres [roughly $^{5}/_{32}$ to $^{5}/_{16}$ inch] in diameter) and of nest-building behaviour in the trouts and salmons.

A large egg with a large yolk to supply food to the developing embryo allows for direct development—that is, the young hatch in an advanced stage, resembling a miniature adult. In more benign environments, such as lakes and the ocean, most salmoniform fishes produce smaller but more numerous eggs, and hatching takes place when the larvae are only partially developed. In many species the larvae are quite unlike the adult form and undergo a rather striking transformation (metamorphosis). Eggs of all freshwater spawning salmoniform fishes are heavier than water (demersal eggs) and develop on or in the bottom of a stream or lake. Marine species typically have pelagic (free drifting) eggs and larvae; the eggs are of neutral buoyancy and thus drift with the currents in the surface layer of the ocean. The eggs and larvae of many deep-sea salmoniforms have not yet been described, and in some species the eggs and larvae may be associated with the ocean bottom.

As far as known, all salmoniform fishes lay eggs and have external fertilization (oviparous fishes). In several of the deep-sea salmoniform families, in which hermaphroditism is common, it is not known if the species are self-fertilizing.

Some of the deep-sea salmoniforms have luminescent organs, one of the functions of which probably is sexual recognition. In the lantern fishes (Myctophoidei), the light organs are arranged in distinctive patterns that distinguish males and females of a species.

Behaviour and locomotion. Only the freshwater salmoniform fishes can be studied in any detail by direct observation. Most of what is known about the deep-sea species is based on preserved specimens, and, for most species, behaviour and locomotion can only be surmised from an examination of the morphology and anatomy.

The generalized body form of trout and salmon is characteristic of active, swift-moving fishes. A trim, fusiform body, powerful caudal (tail) muscles, and a well-developed tail combine to propel the fish against strong currents with a minimum of resistance. These features also give the trout or salmon the ability to leap barrier falls as high as three metres (10 feet) or more.

Adaptations of predators

Predatory fishes that dart out to grasp their prey are exemplified by the pike, in which the dorsal fin is situated posteriorly on the body to act more as a rudder than a keel. The pikelike body form has been evolved independently many times among predatory fishes such as the barracuda (*Sphyraena sphyraena*, of the order Perciformes). Among the deep-sea salmoniforms, however, certain predatory species are sedentary and have only weak swimming ability. Such fish remain immobile until unsuspecting prey ventures close enough to be grasped. Some evidently use a luminous lure to attract their prey.

A peculiar type of locomotion is encountered among the barracudinas (family Paralepididae), marine salmoniforms of the suborder Myctophoidei. The barracudina swims in a vertical plane, darting up and down with the head oriented downward.

The behaviour of a fish toward other members of its species can be highly variable. Often, predator species are territorial and aggressive, whereas plankton-feeding species typically form schools and do not function normally unless they are close to other members of their species. Although behaviour patterns are largely innate and species-specific, striking differences occur between closely related species. Pink salmon fry, on hatching, seek each other and form schools prior to seaward migration. The young of the coho, or silver salmon (*Onchorhynchus kisutch*), however, establish territories and aggressively attack other young cohos that invade their territory. This difference in aggressive behaviour is associated with the longer period of freshwater life and limited food supply experienced by the coho salmon.

Homing by trout and salmon

One fascinating aspect of the behaviour of trout and salmon is their homing instinct—*i.e.*, the ability to return to the stream of their birth after migrating thousands of kilometres in the ocean for one to three years. Homing to the site of birth for reproduction is apparently a rather universal trait among the Salmonidae. Trout, char, and whitefishes in lakes segregate into discrete populations during the spawning season, each at a specific site.

It is now generally accepted that the sense of smell plays the major role in guiding an anadromous trout or salmon to its precise natal stream once it enters a river drainage from the ocean. How it finds the mouth of the river system leading to the natal stream from the open ocean is not yet understood; celestial navigation and detection of fields of gravity by some unknown means have been hypothesized. Several senses besides smell may be used to locate the natal stream. Cutthroat trout (*Salmo clarki*) in Yellowstone Lake, Wyoming, have been found to be able to return to their spawning stream after experimental blocking of the senses of smell and sight.

Homing behaviour has allowed the development of discrete populations among anadromous species of salmon and trout. Different life-history characteristics can be maintained because different populations segregate for spawning, and individuals of a population spawn only with each other, perpetuating hereditary traits. In major river systems such as the Columbia and Fraser in North America, one species may include several distinct races, each having different life cycles; such a situation greatly complicates the management of a species.

Ecology. As with other aspects of the biology of salmoniform fishes, the ecology of species of the family Salmonidae is best known. All species of salmonid fishes evolved in clear, cold water, and they thus require pure, well-oxygenated, cold water; for this reason salmonid fishes are the first species to suffer when water quality is degraded. Other freshwater salmoniforms, although not quite so sensitive to water quality as the salmonid fishes, are also susceptible to the inimical effects of man-induced environmental degradation.

Most salmoniform fishes are predators, feeding on other fish and large invertebrates. The process of evolution, however, works to modify and adapt species for certain ecological specializations in order to exploit a variety of food resources. In the lakes of the Northern Hemisphere, several whitefish species (*Coregonus*) are comparable, ecologically, to the herrings in the ocean. Such whitefishes, which are often called freshwater herrings, cruise the open water of lakes, filtering out minute organisms by straining the water through a fine mesh of gill rakers—minute bony elements attached to the gill arches. The sheefish, or inconnu (*Stenodus leucichthys*), a large, predatory whitefish of the Arctic, demonstrates that evolution for ecological adaptation is occasionally reversible: the adults feed on other fish and have evolved a pikelike body shape and large, powerful jaws, the development of teeth taking precedence over that of the gill rakers; the sheefish is quite unlike the typical whitefish from which it has evolved.

There probably has been strong selection for freshwater salmoniforms to utilize the marine environment for feeding. All groups except the esocoid fishes (pike family and related groups) have species that migrate to the ocean for feeding. This presents a problem of osmotic regulation in waters of different salinities. The physiology of most fishes is fixed for life in fresh water or in the sea, but most of the freshwater salmoniforms are able to live in the sea because they can excrete excess salts through cells in the gills. They also possess well-developed kidneys, which, in the freshwater environment, handle the excess of water that diffuses into their blood via the gills.

Little is known of the ecology of the wholly marine salmoniforms. They may be ecologically grouped by the depths that they inhabit and by their feeding preference. Those found in the twilight zone of the ocean (200–1,000 metres [650–3,300 feet]) consist of plankton feeders and predators. The plankton feeders typically are more active and have a more fully developed and functional swim bladder than is typical of the predatory forms.

Abundance of deep-sea species

Because virtually all primary food production in the oceans takes place in the upper, sunlit layer, the deep-sea fishes live in a food-poor environment. At first, it may seem contradictory that they are able to maintain such numerical abundance; certain features of the biology of the deep-sea salmoniforms, however, allow them to attain great numbers. The body of the typical oceanic salmoniform is feebly developed, appearing to consist of little more than gelatinous material. The skeleton and muscles are reduced, so that little energy is needed to maintain the body. Many of the deep-sea species make nightly migrations to the food-rich surface zone for feeding. The species inhabiting the deepest parts of the ocean must depend on a food supply that filters down from above. This food is concentrated in a narrow bottom layer (the benthic zone), with the result that the benthic species may attain a relatively high abundance.

FORM AND FUNCTION

Features of the generalized salmoniform. *External characteristics.* The tremendous range of structural diversity found in salmoniform fishes has already been mentioned. Comparisons of some of the extreme morphological and physiological modifications with a generalized, standard type can be useful in understanding the evolutionary trends leading to certain specializations. A trout of the genus *Salmo,* such as a rainbow trout or brown trout, can serve as a "standard" for the form and function of salmoniform fishes. The nonspecialized morphology and physiology of a typical trout species allow it to utilize diverse ecological niches during its life. A trout's diet consists of a variety of organisms, and its habitat may vary from small streams, large rivers, or lakes to the ocean. The body and fins are streamlined and symmetrical; the body is covered with small, smooth (cycloid) scales; the fins are formed from soft supporting rays, without spines. A small, fleshy adipose fin is located between the dorsal fin and the tail. The dorsal fin is located midway along the body on the dorsal surface. On the ventral surface, the paired pectoral fins are directly posterior to the head, the paired pelvic (or ventral) fins are directly beneath the dorsal fin, and the single anal fin is positioned beneath the adipose fin. The well-developed tail (caudal fin) connotes a powerful swimming ability. The presence, absence, rearrangement in position, and modifications in size, shape, and function of the various fins are characteristic of the numerous families of Salmoniformes.

Digestive system. The structures associated with feeding and digestion denote the diversity in a trout's diet. The mouth is fairly large with moderate development of nonspecialized teeth on the jaws and on several bones within the mouth. An adult trout can capture and consume a fish about one-quarter its own length without undue difficulty. Feeding on invertebrate organisms, as small as a few millimetres (perhaps $^1/_4$ inch) in length, is facilitated by the gill rakers on the surface of the gill arches; they strain small organisms from a stream of water passing over the gills and funnel them to the esophagus. The well-defined muscular stomach opens by a valve into the intestine. A series of fingerlike appendages opens off of the intestine immediately posterior to the stomach. These appendages, called pyloric ceca, secrete enzymes and provide additional digestive areas to the intestine. Among closely related species of the family Salmonidae, there is a tendency for the more predacious species to have more numerous pyloric ceca. Generalizations relating pyloric cecal development to diet cannot be extended, however, to other fishes. The highly predacious pikes of the genus *Esox* completely lack pyloric ceca, whereas the algae-eating ayu (*Plecoglossus altivelis,* family Plecoglossidae) probably has more numerous ceca than any other fish, up to 400 or more.

Sense organs. Because vision is important in the life of a trout, the eyes are well developed; the retina possesses both rods (for vision in dim light) and cones (for perceiving more acute images and for colour vision). The sense of smell is also highly developed.

The lateral line system

The lateral line nervous system functions as a pressure receptor and a direction finder for objects that move, such as another fish. The lateral line might be considered as a remote sense of touch; it does not, however, function in hearing low-frequency sound waves as was once believed. It has been demonstrated that sound waves are well below the threshold necessary to stimulate the lateral line cells. In trout, the lateral line consists of a series of connected sensory cells (neuromasts) with tiny, hairlike projections. These cells are embedded under the scales along the midline of the body and open to the surface through pores in the scales. An extension of the lateral line system on the head consists of a ramification of sensory canals. In some deep-sea salmoniforms living in the absence of the effects of sunlight, other senses are needed to compensate for vision in perceiving the environment, and the neuromast sensory cells may be exposed on raised papillae, thus increasing their sensitivity.

The swim bladder (or air bladder) has a hydrostatic function, adjusting internal pressure to maintain a weightless condition of neutral buoyancy at various depths. The trouts have a primitive type of swim bladder with a connecting duct from the bladder to the esophagus. The duct is an evolutionary holdover from an ancestor in which the swim bladder was mainly an accessory respiratory organ. Many salmoniform fishes lack the duct, and several deep-sea marine species lack a swim bladder altogether.

Departures from the generalized body plan. From the primitive body plan exemplified by the trouts, it is possible to derive all of the specialized body types of other salmoniform fishes by the elimination of some structures and by the modification, exaggeration, and rearrangement of others.

The pike is an example of a specialized predator whose diet, after the first year of life, consists almost entirely of other fishes. Its success depends on how effectively it captures and consumes other fishes, and its whole morphology and physiology are directed toward this end. A pike has an elongated body with a large head and large, powerful jaws. Its mouth is armed with large, canine-like teeth that can handle large prey. Patches of teeth on the gill arches replace the typical gill rakers. Vision is the primary sense used by pike to detect and capture prey. The visual centre of the brain (optic lobe) is more highly developed than are the centres of smell (olfactory lobes). The eyes have a high proportion of cones to rods in their retinas and are positioned to provide partial binocular vision (*i.e.,* the eyes are aimed in the same direction), sighting down grooves on the snout to aim at moving prey. The body form and position of the fins are specialized for swift, darting movements. The dorsal fin is placed posteriorly, over the anal fin, and, as is typical of other salmoniform fishes with posteriorly oriented dorsal fins, the adipose fin is absent.

The most extraordinary modifications in the basic salmoniform body plan are found among the marine species of the middepths and great depths of the ocean. The more striking adaptations include luminous organs, eyes specialized to function in dim light, feeding adaptations allowing some predatory species to kill and eat a fish as large as themselves, and drastic departures in body shape and fin development.

Luminescent organs

Bioluminescence, the production of chemical light by living organisms, is widespread in nature. Among vertebrate animals, only marine fishes have light organs. Light organs (or photophores) are encountered in many diverse groups of fish. These structures apparently have been evolved independently several times in different groups of fish. It is believed that light organs of fish have evolved from mucous cells of the skin. Salmoniform fishes, particularly species in the suborders Stomiatoidei and Myctophoidei, have developed some elaborate and highly complex light-producing systems. Some structures have lenses, reflectors, and eyelid-like shades. In addition to light cells on the sides of the body, luminous tissue may be found on the head, around the eyes, on fin rays and barbels, and on the ventral surface; in the myctophoid family Paralepididae, an internal duct makes the whole fish glow. The great diversity in the type and position of light organs suggests that they must serve different functions in the groups possessing them. In the family Searsiidae (suborder Alepocephaloidei), a large sac on the shoulder emits a display resembling a shower of sparks when the fish is disturbed. Such a structure probably is a defensive mechanism. Light organs on the head may help in locating food, and those on elongated dorsal fin rays or chin barbels may lure prey. Sexual recognition and territorial behaviour are other suggested functions. Although not all marine salmoniform fishes have light organs, the latter are typically found in species that spend most of their life in the lower twilight and upper dark zones of the ocean. The effects of sunlight essentially disappear at about 700 metres (2,300 feet); the maximum abundance of luminous fishes occurs at about 800 metres (2,600 feet).

Eye adaptations in deep-sea forms

There are some parallels in the development of eyes and light organs in fishes correlated with the depth at which the species lives. Perhaps the most sensitive of all vertebrate eyes is found in fishes inhabiting the dim twilight zone of the ocean; the eyes, specialized to function at very low light intensity, may be greatly enlarged. The retina typically consists entirely of rods with golden pigment to increase sensitivity to blue light of the light spectrum (the last part of the visual light spectrum to be filtered out in water). Another adaptation found in some marine salmoniforms for concentrating weak light is tubular eyes. Fish with tubular eyes appear to be wearing exaggerated goggles. Tubular eyes are aimed in the same direction (binocular vision) and may be directed straight ahead or directly upward. Two sets of retinas are associated with tubular eyes, one on the side of the shaft and one in the normal position at the base. The two sets of retinas function to enlarge the field of vision. A most unusual modification of the eyes is found in the myctophoid genus *Ipnops* (family Ipnopidae), which appears to be eyeless; however, a thin, transparent bony plate on top of the head covers a mass of retinal cells. Evidently such an eye functions to perceive faint luminescence at great depths. Larval stages of a few salmoniforms have eyes extended out from the body on stalks, which are resorbed when the eyes assume a normal position during metamorphosis.

Some grotesque fishes are found among the predatory stomiatoids and myctophoids. The teeth may be developed into tremendously enlarged fangs, which may be likened to daggers, spears, or sabres. The gape of the jaws is sufficiently large to engulf a prey as large as the predator. Stomiatoid predators have a peculiar modification of the anterior vertebral column, which remains unossified, resulting in a flexible jointlike mechanism allowing the head to snap back and enlarge the gape. To allow the swallowing of large prey, the body is soft, distensible, and usually lacks scales; the stomach is highly elastic.

EVOLUTION AND CLASSIFICATION

Evolutionarily important taxonomic characters. Studies of the skeletal system (osteology) and comparative anatomy have produced most of the information used in the classification of salmoniform fishes. At present, however, the taxonomy of Salmoniformes is not well defined because no character or group of characters is exclusive to salmoniform fishes, and little consistent difference in characters occurs among the suborders.

The fishes grouped in Salmoniformes possess a mosaic of primitive characters from which it is possible to derive most of the more advanced orders of teleostean fishes. Both evaluation of the hypothetical evolutionary branching sequences to denote relationships and judgment concerning the primitive or derived (advanced) state of a character are based on an evolutionary principle that a structure lost or highly modified during evolution will never be re-evolved in its original condition; for example, the adipose fin, the mesocoracoid bone of the pectoral skeleton, and teeth on the maxillary bone of the jaw are considered to be primitive salmoniform characters. The absence of these characters represents an advancement; evolutionary lines that have lost these features therefore could not have been ancestral to fishes that possess one or more of these characters. No family of Salmoniformes has all of the primitive characters, but the families Salmonidae and Osmeridae have most of them.

Skeletal features

The primitive dentition pattern of the ancestral salmoniform would be with teeth on the premaxilla and maxilla (bones of the upper jaw) and on the dentary bone of the lower jaw, with the maxilla clearly dominant over the premaxilla in forming the gape of the upper jaw. Inside the mouth, on the upper surface, teeth would be on palatine and pterygoid bones on each side and on the median vomerine bone. On the lower surface of the mouth, teeth would cover the tongue and occur on a plate overlying the basibranchial bones (between the gill arches). Many separate lines of salmoniforms show evolutionary advancement for the loss of teeth and the dominance of the premaxilla over the maxilla, with loss of teeth on the maxilla.

The primitive structure of the pectoral girdle, associated with the ventral position of the pectoral fin, consists of an additional supporting bone—the mesocoracoid. The advanced condition, related to a more dorsally positioned pectoral fin, is the loss of the mesocoracoid.

The primitive salmoniform condition of the caudal skeleton has three separate vertebral centra (the centrum is the main body of the vertebra) and associated bony elements functioning in support of the bony plates (hypural plates), which form the base of the tail. Such a structure is a vestige of the upturned tail (heterocercal tail) characteristic of a more primitive stage of fish evolution. Other such vestiges found in some salmoniform fishes include abdominal pores (minute ducts from the body cavity to the exterior), remnants of a preteleostean type of intestine (spiral valve), and the absence of oviducts in females. Evolution in several salmoniform lines has reduced by fusion the three supporting caudal vertebrae to two or, more commonly, one. The primitive type of salmoniform swim bladder is connected to the esophagus by a duct. The absence of the duct or absence of the swim bladder is the advanced state. The primitive salmoniform had pyloric ceca on the intestine—a variable trait among living species. The light organs of some marine salmoniforms are an advanced character and are considered to be derived from mucous cells.

Annotated classification. The classification presented here is based on that of P.H. Greenwood *et al.*, with some modifications incorporated from more recent publications. These alterations consist of transferring the family Salangidae from the suborder Galaxioidei to the suborder Salmonoidei, placing the family Bathylaconidae in the suborder Alepocephaloidei (eliminating the suborder Bathylaconoidei), and the recognition of two additional families: Prototroctidae (in the suborder Galaxioidei) and Searsiidae (in the suborder Alepocephaloidei).

ORDER SALMONIFORMES

A diverse group of fishes with a mosaic of primitive characters. Typically fusiform or elongated predatory fishes. Adipose fin usually present; dorsal fin and pelvic fins typically placed midway along body; fin rays without true spines; pectoral fin generally in ventral position. Scales, if present, typically smooth (cycloid). Light organs present in several marine families. Caudal skeleton with 1 to 3 vertebral centra functioning in support of tail. Fossils from Cretaceous.

Suborder Salmonoidei

About 100 species; 10–150 cm (4 to 60 in.) long; freshwater, anadromous, or marine; Northern Hemisphere. Adipose present in all species; swim bladder with open duct; maxilla dominant over premaxilla in upper jaw; no light organs; intestine with pyloric ceca (except Salangidae); tail support on 3 distinct vertebral centra in Salmonidae, fused into single element in other families. Suborder includes the families Salmonidae (including Coregonidae and Thymallidae), salmons, trouts, chars; Osmeridae, smelts; Plecoglossidae, ayu; and Salangidae, icefishes.

Suborder Galaxioidei

About 50 species; 7.5–40 cm (3 to 15¾ in.) long; freshwater, anadromous, or catadromous; Southern Hemisphere. Adipose fin absent in Galaxiidae, present in other families; swim bladder with or without duct; relationship of maxilla and premaxilla variable among genera. Light organs absent. Pyloric ceca present or absent. Tail support on 1 or 2 vertebral centra; mesocoracoid bone of pectoral girdle absent; teeth present on mesopterygoid bone in roof of mouth. Suborder contains the families Galaxiidae, no group name; Retropinnidae, New Zealand smelts; Aplochitonidae, South American trouts; and Prototroctidae, New Zealand "grayling."

Suborder Esocoidei

Ten species; 5–150 cm (2 to 60 in.) long; freshwater; Northern Hemisphere. Adipose fin lacking; swim bladder with open duct; maxilla without teeth; pyloric ceca lacking; pectoral girdle without mesocoracoid bone; tail support on 3 separate vertebral centra; 2 sets of paired ethmoid bones on snout region of skull. Suborder includes the families Esocidae, pikes; and Umbridae (including Dalliidae), mudminnows.

Suborder Argentinoidei

About 50 species; 3–40 cm (about 1 to 15¾ in.) long; marine; worldwide. Adipose fin present on most species; swim bladder without duct or absent entirely; maxilla and premaxilla reduced, without teeth; light organs present in several species; tail support on 2 vertebral centra. Suborder includes the families Argentinidae (including Xenophthalmichthidae and Microstomatidae), argentines; Bathylagidae, deep-sea smelts; and Opisthoproctidae (including Dolichopterygidae, Macropinnidae, Winteridae), barreleyes.

Suborder Stomiatoidei

About 350 species; 2.5–45 cm (1 to 17¾ in.) long; marine; worldwide. Adipose fin present or absent, some species with both a dorsal and a ventral adipose fin; swim bladder without duct or absent entirely; maxilla the dominant bone of the upper jaw; some species with greatly enlarged, depressable teeth; anterior vertebrae sometimes unossified; light organs present in most families; members of some families with chin barbel, which may be a highly elaborate structure; tail support on single vertebral centrum. Suborder includes families Stomiatidae, scaly dragonfishes; Gonostomatidae, bristlemouths; Sternoptychidae, hatchetfishes; Astronesthidae, snaggletoothed fishes; Melanostomiatidae, scaleless dragonfishes; Chauliodontidae, viperfishes; and Idiacanthidae, black dragonfishes.

Suborder Alepocephaloidei

About 120 species; 3–700 cm (about 1 in. to about 23 ft); marine, deep-sea; worldwide. Adipose fin lacking; swim bladder lacking; teeth small; intestine with pyloric ceca. Light organs present in some species (on raised papillae). Tail supported by 3 vertebral centra. Suborder contains the families Alepocephalidae, smoothheads; Bathylaconidae, bony throats; Searsiidae, tubeshoulders; and Bathyprionidae.

Suborder Myctophoidei

About 470 species; 2.5–200 cm (1 in. to 6½ ft) long; marine; worldwide. Adipose fin usually present; swim bladder present in some species, if present, with duct. Premaxilla dominant over maxilla in upper jaw; no mesocoracoid bone in pectoral girdle. Light organs present in all species of family Myctophidae, but absent in species of most other families; tail support on single vertebral centrum. Suborder includes the following 15 families: Myctophidae, lantern fishes; Aulopodidae, thread-sail fishes; Synodontidae, lizard fishes; Harpadontidae, bombay duck; Chlorophthalmidae, green-eyes; Bathypteroidae, spider fishes; Ipnopidae, grid-eye fishes; Paralepididae, barracudinas; Omosudidae, hammerjaw; Alepisauridae, lancet fishes; Anotopteridae, javelin fish; Evermannellidae, sabre-toothed fishes; Scopelosauridae, paperbones; Neoscopelidae, blackchin; Scopelarchidae, pearleyes.

Alternate views on the nearest relatives of the salmoniforms

Critical appraisal. Previous schemes of fish classification have been based mainly on the work of the British ichthyologist C.T. Regan and the Soviet ichthyologist L.S. Berg. Both Regan and Berg grouped most of the generally primitive fishes with soft fin rays and smooth scales in an order with the herring family, Clupeidae. Regan called this order Isospondyli, and Berg used the name Clupeiformes. Such a classification considered Isospondyli or Clupeiformes as the most primitive of the teleostean fishes and as ancestral to all other advanced orders of Teleostei. The work of Greenwood and his colleagues clearly demonstrated that the classifications of Regan and Berg were without evolutionary reality, that the fishes classified as

Clupeiformes or Isospondyli, as formerly arranged, were not all derived from a common ancestor but consisted of several unrelated groups. The true herrings (family Clupeidae and their direct derivatives) possess some unique characters, such as the structures involved with the connection of the swim bladder to the inner ear, not found in any other teleostean fishes; the herrings thus are not very likely to have been the progenitors of all other modern teleosts. The order Salmoniformes was created to remove several diverse groups of dubious relationships from the order Clupeiformes; these groups are thus considered as the basal stocks in the evolutionary radiation of teleostean fishes. The order Iniomi of Regan (Scopeliformes of Berg) was placed as a suborder, Myctophoidei, in Salmoniformes. It should be emphasized, however, that this taxonomic revision has added little new knowledge concerning the relationships among the various suborders grouped in Salmoniformes. At present no coherent picture of evolutionary affinities among the suborders and with other orders has emerged. Undoubtedly the present interpretation of Salmoniformes will undergo major revisions in the future both in structure and in its implications regarding the evolutionary links leading to other teleostean orders. Publications by D.E. Rosen and by the British ichthyologist Colin Patterson consider new evidence and a more critical interpretation of previous data; they have separated the suborder Myctophoidei from the Salmoniformes into the order Myctophiformes and have suggested that the suborder Giganturoidei of the order Cetomimiformes should be placed in Salmoniformes. The myctophoid fishes are well separated from other salmoniforms, having undergone their own evolution at least since Cretaceous times (about 100,000,000 years ago)—fossil records of four families are known from Cretaceous deposits—and recognition of the order Myctophiformes seems justified.

Hopefully, new evidence will be forthcoming from future studies providing new insights into the evolutionary events that transpired from 50,000,000 to 100,000,000 years ago and resulted in the evolution of the various families and suborders of Salmoniformes. From such evidence, a better understanding of the ancient divergences leading to the bulk of the present teleostean fishes should follow. Ichthyologists may have exploited anatomical characters to the limit of potential information yield, but it is to be expected that new fossil finds of extinct, intermediate groups will yield valuable information.

New techniques for examining chromosomes and comparing their number, size, shape, and content and for comparing the structure of certain evolutionally stable protein molecules are promising approaches for the interpretation of evolution. (R.J.B.)

Characins, catfishes, minnows, carps, and allies (Ostariophysi)

The fishes of the superorder Ostariophysi include the majority of freshwater fishes throughout the world. Familiar representatives of this group are the minnows, suckers, characins, loaches, gymnotid "eels," and innumerable catfishes. The 31 recognized families of catfishes constitute the order Siluriformes, the remaining 26 families the order Cypriniformes. Humans consume huge quantities of these fishes for food and derive pleasure from the beauty of tropical aquarium fishes. A few harmful species can inflict painful injuries; some others serve as intermediate hosts for parasites of humans. Strange and fascinating behaviour is exhibited by many of these fishes—nest building, oral incubation, egg laying in mollusk shells, walking and flying, air breathing, production of sound and electricity, and communication by chemical secretions.

GENERAL FEATURES

Size range and diversity of structure. Most ostariophysians are small to moderate in size, from two to 30 centimetres (about one to 12 inches) long; others rank among the giants of the freshwater world. The elegant mahseer (Cyprinidae) of Asia grows to two metres ($6^1/_2$ feet) long and weighs 90 kilograms (200 pounds); the wels, a Eurasian catfish (Siluridae), attains a length of 4.5 metres (15 feet) and a weight of 300 kilograms (660 pounds). The extent of morphological diversity is at least as great as that in any other group of living vertebrates.

Ostariophysians abound in nearly all freshwater habitats, including subterranean caverns and those on all major landmasses and continental islands of the world except for Greenland and Antarctica. A few invade brackish waters, and two families consist largely of marine species. Approximately 6,000 species are recognized, about one-fourth of all known species of fish. Their undisputed success may be attributed at least in part to two remarkable features: a sense of hearing more acute than that in any other group of fish and a warning system by chemical communication unique among fishes.

Drawing by D.P. Janson

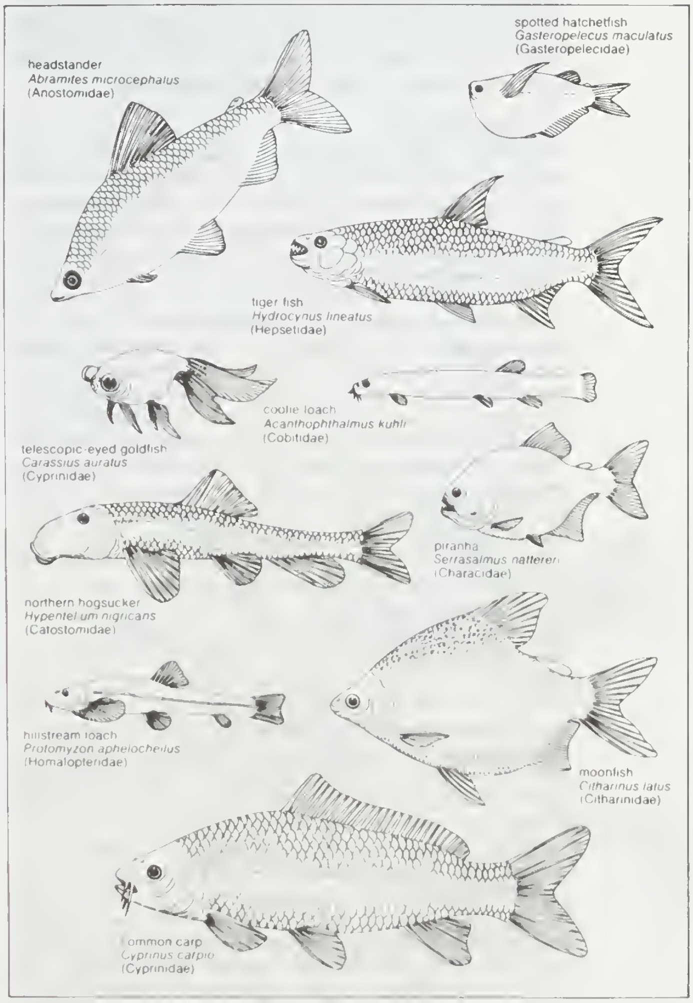


Figure 25: Body plans of representative ostariophysian fishes.

Importance. Many moderate to large ostariophysians are utilized for food, and commercial fisheries harvest huge quantities of marketable species. The common carp (*Cyprinus carpio*), originally from China, has been introduced nearly worldwide and is extensively cultured in the warmer regions. Other Chinese carp under cultivation include the grass carp (*Ctenopharyngodon*), silver carp (*Hypothalmichthys*), snail carp (*Mylopharyngodon*), and bighead carp (*Aristichthys*). Culture of the channel catfish (*Ictalurus punctatus*) is an important industry in the southern United States. Numerous ostariophysians provide sport fishermen with recreation and food; several rank among the world's prized game fish; *e.g.*, mahseers (several species of *Barbus*) of Asia and the dorado (*Salminus maxillosus*) of South America.

Among the most popular aquarium specimens are the characins, tetras, rasboras, danios, barbs, loaches, and innumerable catfishes. Adaptability of many ostariophysians to aquarium life has resulted in their widespread use as experimental animals in scientific research. Foremost among these are the goldfish (*Carassius auratus*) and the common carp.

In eastern Asia and parts of Europe,.humans frequently become infected with liver flukes acquired by eating raw or imperfectly cooked fish. The carps, especially *Ctenopharyngodon idellus,* are the second intermediate host of the Chinese liver fluke (*Clonorchis sinensis*). Many cyprinids serve as intermediate hosts for the cat liver fluke (*Opisthorchis felineus*). Domestic animals similarly become infected with flukes and tapeworms.

Some ostariophysians are pests or are potentially dangerous to man. The common carp is a nuisance in many localities in the United States. Introduced species such as the walking catfish (*Clarias batrachus*) pose a serious threat to the native fauna. In South America, on occasion, the piranha (*Serrasalmus*) voraciously attacks man and domestic animals, and the diminutive candiru (*Vandellia cirrhosa*) can penetrate the urogenital openings of human bathers and cause intense pain and hemorrhaging.

NATURAL HISTORY

Behaviour. *Reproductive cycle.* Like most fishes, ostariophysians are bisexual; eggs develop in the ovaries of the female and spermatozoa (milt) in the testes of the male. In temperate zones most species breed in the spring, when water temperatures are rising and day lengths (photoperiods) are increasing. In tropical regions many fishes spawn the year round. All ostariophysians lay eggs; none give birth to living young. Eggs are fertilized externally in the water in all species except a few auchenipterid catfishes and the South American characins of the subfamily Glandulocaudinae in which fertilization is internal. Development is direct; newly hatched individuals do not pass through morphological changes (metamorphosis) but instead are miniature replicas of their parents. The age at sexual maturity depends on the species and relative body size. Many small species reproduce when only a few months old and rarely live more than one to a few years. Large species attain sexual maturity when several years old, and, in captivity, the common carp has been known to live more than 50 years.

Breeding. Distinct pairing occurs in most ostariophysians, and courtship behaviour in characoids and cyprinoids often consists of elaborate displays by males in brilliant nuptial coloration. The eggs are heavier than water (demersal) and sink; most are sticky and adhere to the surface or to various objects. Characins and cyprinids generally deposit their eggs among aquatic plants, under stones and logs, or in shallow pits in gravel and sand. Among the many exceptions is the characin *Copeina arnoldi;* the female actually leaps out of the water to lay her eggs on the undersides of overhanging leaves (or, in captivity, of aquarium covers), to which she clings, joined by the male, during egg deposition. The parents then splash water on the eggs during development. The female bitterling (*Rhodeus sericeus*) deposits its eggs in the gill cavity of freshwater mussels by means of an elongated ovipositor, which she inserts into the mussel's incurrent siphon. Catfishes choose breeding sites in streams and ponds, generally in quiet water among plants or on mud, sand, gravel, or debris. The nest may be a simple circular depression (as in bullheads) or a tunnel-like affair in the bank (as in the channel catfish). Migrations comparable to those of salmon and eels are unknown among the ostariophysians, but the tendency to migrate occurs among suckers (Catostomidae), which swarm upstream into small tributaries and spawn over gravel or sand bottoms, and in other riverine species, as the mahseer (*Barbus tor*) and the African tiger fish (*Hydrocynus*).

Egg laying on leaves out of water

Parental care. Although many species exhibit no parental care, nest building and egg guarding are widespread among this group of fishes. Some cyprinids, such as the chubs (*Nocomis*), build massive pyramidal nests of stones; they desert the nests once spawning is completed. Other species of breeding minnows often swarm over these nests, and hybrids frequently are produced by the mixing of eggs and sperm from different species. The eggs of characids are commonly guarded by the male. Catfishes provide their eggs with considerable protection, either by guarding nests or by carrying the eggs with them. Oral incubation is practiced in sea catfishes (Ariidae); the male carries from 10 to 50 marble-sized eggs in the mouth cavity until hatching. The male continues to protect the hatchlings in his mouth even after the young have begun to feed independently. In certain species of banjo catfishes (Aspredinidae), the eggs are anchored to spongy tentacles on the underside of the female's abdomen. Some female callichthyid catfish carry eggs on the abdomen only for fertilization; others deposit their adhesive eggs in froth nests and guard them. The loricariid catfishes emply various methods; some lay adhesive eggs in cavities, others carry them under the lower lip, and a few deposit them on rocks, where they are cleaned, fanned, and guarded by the male.

Defense. Ostariophysians with bright colours and gaudy patterns are popular among tropical-fish fanciers; however, many other small species are somberly coloured, relying on this protective coloration for passive defense from enemies. Large carnivorous forms such as the African tiger fish and the South American piranhas have powerful jaws and strong teeth, extremely effective weapons for defense as well as for offense. Most catfishes and some Old World cyprinids possess spines (hardened fin rays) in the dorsal and pectoral fins. The spines alone afford a considerable degree of protection; in addition, venom glands develop at the base of the spines in some bullheads and madtoms of North America (Ictaluridae), labyrinthic catfishes (Clariidae), and sea catfishes (Ariidae and Plotosidae). Painful but rarely fatal injuries result when the skin of a victim is punctured and venom injected.

Protection by poisonous spines

Although a variety of freshwater fishes can generate an electrical charge, only two develop sufficient voltage to stun other animals, including man—the electric eel (*Electrophorus electricus*) and the electric catfish (*Malapterurus electricus*).

Ecology. *Habitat and distribution.* Ostariophysians are the dominant fishes in virtually all types of freshwater habitats throughout the tropical, temperate, and subarctic regions of the world. Only a few species of the families Cyprinidae and Aspredinidae are known to invade low-saline or brackish waters. The only truly marine members of this superorder are the sea catfishes (Ariidae and Plotosidae), which inhabit tropical coasts. Some plotosids, however, live in freshwater.

The upper regions of small mountain streams are characterized by steep gradients, waterfalls and rapids, and torrential currents. Here occurs a variety of ostariophysians (Homalopteridae, Sisoridae, Akysidae, Loricariidae, Astroblepidae), which exhibit fascinating structural adaptations, such as holdfast organs and specialized respiratory mechanisms. In river systems where the gradients are not steep, currents are slow and quiet pools alternate with riffles, large numbers of characins, cyprinids, and suckers and other types of catfish are conspicuous elements in the fauna. In the sluggish waters of large rivers live large species of suckers, cyprinids (*e.g.,* carp), and many catfishes generally characterized by environmental tolerances and nonrestrictive feeding habits. Ponds and lakes also support large populations of characids, cyprinids, catostomids, and siluroids that prefer and are adapted to standing-water habitats. Although a few are benthic (bottom-dwelling) forms, most of the characins and cyprinids tend to live and feed in the middle and upper layers of the water column. Suckers, loaches, and most catfishes are typically ethic animals and thus are highly adapted to such an existence. Catfishes are generally most active at night or under conditions of reduced light intensities.

Cave fishes

Among the most unusual habitats for fishes are those in subterranean waters, wells, and caves. A relatively large number of ostariophysians, belonging to unrelated families, present a striking example of convergent adaptation to life in more or less total darkness. The evolutionary trends have led to a reduction or loss of eyes, loss of pigment, and special development of certain sense organs, especially the lateral-line system, to compensate for the loss of sight. Ostariophysians adapted to such a mode of life include six genera of cyprinids in Africa, the Middle East, and Java; a characin (*Astyanax jordani*) from Mexico; ictalurids from the U.S. (*Trogloglanis* and *Satan*) and Mexico (*Prietella*);

six genera of pimelodids and trichomycterids from South America; and two genera of clariids from Africa.

Feeding habits. The remarkable diversity of feeding habits among ostariophysians is associated with a fantastic variety of adaptations in mouth shapes and tooth types (especially in the suborder Characoidei) probably unsurpassed by any other group. At one extreme are certain cyprinids (*e.g., Notropis atherinoides*) with highly developed gill rakers that strain phytoplankton (minute plants) from the water. Mountain-stream fishes (*e.g.,* Gyrinocheilidae, Homalopteridae, Loricariidae) possess suckerlike lips for scraping algae from the rocks; their teeth are minute or entirely lacking. Because they devour large quantities of plants, herbivores such as the Chinese grass carp are used experimentally to control vegetation in weed-choked waters. Omnivores are especially common among the characins and catfishes. Suckers, long-snouted knife fishes, many catfishes, and some minnows suck up mud and bottom debris, extract the nutriments, and eject the residue. Small carnivorous species consume insect larvae, small crustaceans, worms, mollusks, and other invertebrates. At the top of the food chain are the voracious predators, the most famous of which are the piranhas. Although modest in size, they have short, powerful jaws armed with razor-sharp teeth. These fearsome predators often occur in large schools and can quickly strip the flesh from their victims. Other fishes are their usual prey, but cattle and occasionally humans are also attacked. Probably the largest predatory ostariophysian is the tiger fish, which attains a weight exceeding 45 kilograms (approximately 100 pounds); its huge, sharp teeth and large, tunalike tail endow it with ferocity and speed. Parasitic habits are rarely found among bony fishes, but certain species of trichomycterid catfishes attach themselves to the gills of other fishes and feed on their hosts' blood.

Feeding habits of piranhas

FORM AND FUNCTION

Distinguishing characteristics. *Weberian apparatus and swim bladder.* The single character unique to the superorder Ostariophysi is the presence of the so-called Weberian apparatus, a complex connection between the inner ear and gas bladder (swim bladder). It is formed by the modification of the first four (or five) vertebrae immediately behind the skull, small portions of which have become separated and form a chain of four paired bones, or ossicles, named from front to back the claustrum, scaphium, intercalarium, and tripus. The first is in contact with a membranous window, or extension of the inner ear; the last touches the anterior wall of the swim bladder. The diverse modifications of the Weberian apparatus are diagnostic of orders and certain families; *e.g.,* the claustrum is absent in Gymnotidae. Although much remains to be learned about its functions, it is known to serve as a hearing organ. Changes in volume of the swim bladder due to sound waves in the water cause the ossicles to move and transmit pressure changes to the ear.

The gas bladder varies in shape and size but typically consists of two, sometimes three, chambers. In bottom-dwelling fishes, such as the Homalopteridae, Cobitidae, and many catfishes, the posterior chamber is greatly reduced and the anterior one often more or less surrounded by a bony capsule. In some catfishes (Sisoridae), only the anterior chamber remains, and it may be encapsulated with bone.

Body covering. The nature of the body covering is variable. Most cypriniforms possess cycloid scales (smooth, overlapping scales more or less circular in shape). Exceptions are found among the Ctenoluciidae, Distichodontidae, Citharinidae, and Ichthyboridae, which have ciliate, or ctenoid, scales (*i.e.,* posterior margins of scales with fine teeth). Most catfishes have lost the scaly covering and are naked, but several families possess bony plates forming an overlapping armour on the sides of the body (Doradidae, Callichthyidae, Loricariidae).

Bony armour

Fin spines and adipose fin. Ostariophysians possess segmented, branched, flexible, soft rays in the fins, unlike the stiff spines of perchlike fishes. In some species, however, soft-ray elements may fuse during development and give rise to a spinous ray (usually called a spine), commonly found in the dorsal and pectoral fins of most catfishes and in the dorsal and anal fins of some Old World cyprinids. The presence or absence of these spines is frequently diagnostic for genera and families.

An adipose fin consists of a small to elongated fleshy or fatty structure without fin-ray supports, located dorsally between the rayed dorsal fin and caudal (tail) fin. It is present in most ostariophysian fishes.

Barbels. Diverse morphological differences in the mouth region are related to the type of diet and to the modes of locating, capturing, and ingesting food. Barbels are short to filamentous, fleshy, fingerlike projections located at the corners of the mouth or on the snout and chin of many suctorial and bottom-feeding fishes (some minnows, loaches, and catfishes). Barbels are highly sensitive to touch, and they bear numerous taste buds. Taste and touch probably function together in the selection of food before ingestion.

Teeth. Teeth may be present along the jaws, in the roof of the mouth, on the tongue, or in the pharynx, or they may be entirely absent. In the minnows (Cyprinidae) and suckers (Catostomidae), the mouth is toothless, but an array of teeth is borne on a pair of branchial bones, the lower pharyngeals, located in the throat. In the minnows the pharyngeal teeth, arranged in one, two, or three rows, press or bite against a horny pad in the roof of the mouth. They have undergone specialization paralleling the diversity found in jaw teeth of other fishes. Vegetarians such as the carp have grinding, molar-like teeth; carnivores have pointed or hooked teeth. Suckers have numerous pharyngeal teeth aligned in a single row. Oral and pharyngeal teeth are of great value in classifying many families of ostariophysians.

Secondary sexual characteristics. With the onset of the breeding season, many secondary sexual characteristics

Drawing by D P Jenson

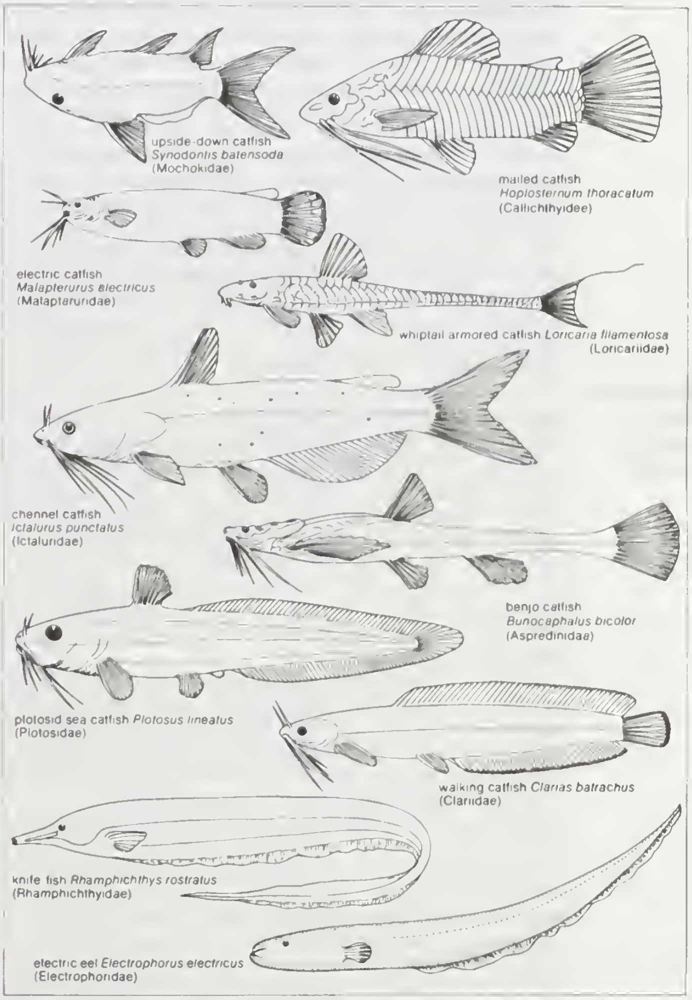


Figure 26: Body plans of representative ostariophysian fishes.

Sexual coloration

develop: size differences, nuptial coloration, enlarged and modified fins, breeding tubercles, and contact organs. These features are related chiefly to courtship and mating, but differences in size obviously play a role in guarding nests and care of the young; the sex that exercises parental care is usually the larger. Brilliant red, orange, yellow, green, and blue coloration may develop on various parts of the head, body, and fins, especially in the males. Some characids and cyprinids are among the most beautiful of all fishes. The male usually has larger and more brightly coloured fins than the female. In some characins, the median and pelvic fins of the males may possess small hooks or contact organs, which aid in maintaining contact with the female during spawning. In the six families of cyprinoids, breeding tubercles, or pearl organs (epidermal excrescences), develop on the head, body, and fins of males under the influence of sex hormones. The tubercles function in maintenance of body contact during spawning, in defense of nests and territories, and possibly in the stimulation of females during breeding.

Sexual differences among the siluroids are more marked in the highly specialized families. Pelvic fins of female ariid catfishes and, to a lesser extent, of ictalurid catfishes show specialized developments whose functions are not yet fully known. Some male loricariid catfishes develop elaborate dermal, branching growths and spines around the head; in others, the lower lip is enlarged to accommodate the transport of eggs.

Adaptations for locomotion. *Swimming.* The body of most ostariophysian fishes is more or less streamlined, taking the most efficient form for movement through water. In this highly diversified group, however, a large array of adaptations occurs. Lateral compression (flattened from side to side) is common, especially among characins and cyprinids that inhabit quiet, weedy lakes, ponds, and backwaters. Extreme examples are the flying hatchetfishes (Gasteropelecidae) and the knife fishes (Rhamphichthyidae and Apteronotidae). Depressed body form (flattened from top to bottom), especially in the head region, is widespread among fishes spending much time on or near the bottom or under rocks and similar objects (most catfishes) or among those inhabiting torrential mountain streams (Homalopteridae, some Loricariidae). An elongated, eellike form has evolved in certain loaches (Cobitidae) and electric eels (Electrophoridae), fishes that live on soft, muddy, and sandy bottoms or in rock crevices.

Departures from normal posture

The most common form of locomotion among ostariophysians is swimming by lateral undulations of the body, resulting from the contractions of muscles along the sides of the body and base of the tail. These undulating flexures culminate in a powerful back and forth sweeping of the caudal fin, which produces as much as 85 percent of the total thrust. Some fishes have departed from the normal horizontal swimming posture. The headstanders (Anostomidae) move with the head pointing downward at a slant; some of the pencil fishes (Hemiodontidae) assume a tail-standing position. Most bizarre of all are the upside-down catfishes (Mochokidae) of Africa, which can swim either in the normal position or inverted, with the belly uppermost. In *Synodontis batensoda,* the coloration of the belly is darker than the back, a reversal of the usual pigmentation pattern. Displacement of the gas bladder toward the underside is a further adaptation to this unusual swimming behaviour.

In fishes with specialized modifications of body form and habits, the fins are frequently modified and used for propulsion. The electric eels and knife fishes (Gymnotoidei) have lost the dorsal fin and, in some cases, the caudal fin. Slow forward and backward movements are made possible by undulations of an extremely long anal fin.

Associated with locomotion is the need for maintaining position in the water, particularly in the rapid torrents of mountain streams. A variety of modifications have evolved that function as holdfasts, anchoring the fish to rocks or similar objects. The hill-stream loaches (Homalopteridae) of southeastern Asia possess a large ventral suction disk formed by the expanded pectoral and pelvic fins. Some of the mountain-stream catfishes (Sisoridae) of Asia have an adhesive organ on the thorax (chest). Mountain-inhabiting catfishes of South America may use a suckerlike mouth (Loricariidae) or employ a combination of a disklike mouth and disklike paired fins (Astroblepidae) for adhesion to the surface.

Walking and flying. A few ostariophysians have the capability to emerge from their aquatic abode and move over land, climb walls, or even fly through the air. The walking catfish (*Clarias batrachus*), recently introduced into southern Florida, uses its pectoral-fin spines as anchors to prevent jackknifing as its body musculature produces snakelike movements and can progress remarkable distances over dry land. Using suction disks and fins, the mountain-stream catfishes (Sisoridae and Astroblepidae) can climb vertical rock walls above the water surface.

Flying by characins

The small hatchetfishes, or flying characins (Gasteropelecidae), of South America normally swim near the surface of the water but are capable of jumping clear and flying short distances. They vibrate enlarged pectoral fins rapidly back and forth by using highly specialized musculature on the shoulder girdle.

Air breathing. Although gills are typical respiratory structures in fishes, many freshwater species occupy habitats where the oxygen may be depleted occasionally or where droughts may force them to live out of water temporarily. These fishes have evolved a variety of airbreathing organs, most of which are outgrowths or pouches from the pharynx, branchial (gill) chamber, or digestive tube. Some catfishes (*Clarias* and *Heterobranchus*) of Asia and Africa have treelike respiratory structures extending above the gill chambers; others (*Heteropneustes*) have elongated, tubular, lunglike sacs extending backward as far as the tail. The electric eel is a mouth breather, gaseous exchange taking place through the wrinkled mucous membrane lining the mouth cavity. Some fishes actually swallow air into the lower part of the digestive tract, which then also serves as a respiratory structure. In the armoured catfishes (*Doras, Plecostomus, Callichthys*) of South America the thin-walled stomach serves this function. The loaches swallow air into a reservoir-like bulge from the intestine and void the remaining gases through the anus.

Communication and sensory perception. *Sound.* Sounds produced by ostariophysians are usually associated with the swim bladder. Minnows produce noises by expelling air through the pneumatic duct, which connects the gas bladder with the digestive tract, and the mouth; loaches do the same by expulsion through the anus. In several catfish families the expanded ends of a springlike mechanism (derived from modified portions of the fourth vertebra) are attached to the swim bladder. The contraction of muscles extending from the spring mechanism to the skull cause the springs and bladder wall to vibrate rapidly, producing a growling or humming noise. In other catfishes the rubbing or grating movements of the dorsal and pectoral spines produce sounds.

Hearing range

The sense of hearing in Ostariophysi is more highly developed than in any other fishes. The walls of the gas bladder are set in vibration by waves of underwater sound, and the Weberian ossicles then increase the amplitude of these vibrations, transmitting them to the internal ears. This combination is analogous to that of a hydrophone and endows these fishes with a remarkable sensitivity to sound. The normal frequency range detectable by ostariophysians is from 16 to 7,000 hertz (cycles per second); for some characins the maximum is 10,000 hertz. Among other functions, sound production and hearing in fishes may assist in bringing schooling fishes together; even more significant is the role of sound in reproduction. Experiments with North American cyprinids provide evidence that sounds are produced by both sexes and may serve for sexual recognition. A male is able to distinguish the calls of females of his own species from those of closely related species. Consequently, sounds may serve as isolating mechanisms in maintaining the genetic integrity of the species. For fishes living in muddy waters, sounds may be a vital communication link between individuals, especially in the breeding season. It is reasonable to suggest that the combination of sound production and acute hearing may have aided the ostariophysians in attaining their dominant role in freshwaters.

Production of electric currents

Electric organs. Members of the suborder Gymnotoidei and of the siluriform family Malapteruridae possess the unusual capacity to generate electricity. The best known and most powerful of this group is the electric eel (*Electrophorus electricus*). The electrical organs, three on each side of the body, are derived from modified muscle tissue. The force of the discharge has been measured at 350 to 650 volts and can produce a current strong enough to stun animals as large as a horse or man. The electric catfish (*Malapterurus electricus*) can deliver shocks up to 450 volts, but this power is apparently used only as a defensive measure. The electrical organ of this species, also derived from muscle tissue, consists of a specialized, gelatinous coat of tissue that sheathes most of the body just under the skin.

The gymnotid eels and knife fishes (*Gymnotus* and other genera) produce currents of low voltage only, emitting a continuous series of pulses (from 35 to 1,700 hertz), which create an electrical field around the fish. When this field is broken, either by a moving animal or by inanimate objects in the vicinity, the fish can locate the objects, which otherwise would be difficult to see at night or in muddy water. Experiments indicate that electrical cues may also facilitate social interactions. Perception of electrical stimuli occurs in specialized electrical receptors in the skin, and portions of the brain are enlarged to process electrosensory information.

Taste and smell. Catfishes and other fishes living in muddy waters have relatively poor vision but possess chemosensory acuity. Lips, barbels, and most of the body are covered with innumerable taste buds. Experiments have proved that taste plays a leading role in the location of food by these fishes.

Studies on the sense of smell have isolated odours emanating from mucus produced in the skin, from secretions of the gonads, and from other body parts. These odours, chemical signals called pheromones, provide a means of communication between individuals of the same or different species. Certain minnows (Cyprinidae) can discriminate between the odours of at least 15 species of fishes belonging to eight different families. The social behaviour of bullheads (*Ictalurus*) and other ostariophysians is related to a system of communications utilizing chemical signals. An individual not only recognizes individuals of other species but can identify and remember the identification of a particular individual of its own species after a time lapse of three weeks. Territorial and communal behaviour are evidently influenced by different pheromones.

Reactions to wounded companions

Alarm substances. In 1938 an Austrian biologist, Karl von Frisch, introduced an injured minnow (*Phoxinus*) into a school of the same species and observed that the school rapidly retreated and became very frightened. By experimentation he demonstrated that a chemical substance released from the lacerated skin produced a fright reaction when perceived through the nasal organs of other fishes. This "alarm substance," secreted by specialized cells in the epidermis, is released only when the skin is injured. Alarm substances are present in almost all species of ostariophysians tested (except for a few species of Characidae, Hemiodontidae, Chilodontidae, and Rhamphichthyidae) and are absent in all nonostariophysian fishes examined. Although the fright reaction appears to be important insurance for the individual against predation, the alarm substances are of greatest value among those species exhibiting social behaviour by warning other members of the school. It is possible that alarm substances and the fright reaction have contributed markedly to the biological success of the Ostariophysi.

CLASSIFICATION

Distinguishing taxonomic features. Many characteristics are useful in classifying this large, diverse superorder—the nature of the body covering; presence or absence of barbels, fin spines, and adipose fin; modifications of mouth and fins; types of teeth. Less obvious but especially significant are numerous skull features, specializations of the Weberian apparatus, configuration of the gas bladder, and fusions of vertebral elements.

Annotated classification. This classification, a recent revision by P.H. Greenwood (U.S.) and colleagues, raises to family rank many groups of fish often treated as subfamilies. The smallest families are grouped for brevity or are included under a closely related family.

SUPERORDER OSTARIOPHYSI

Gas bladder and internal ear connected by chain of ossicles (Weberian apparatus). All inhabit freshwater unless otherwise noted.

Order Cypriniformes

Body usually covered with cycloid scales; about 3,500 species.

Suborder Characoidei

Cretaceous (about 136,000,000 years ago to present). Mouth not protractile; jaws toothed. Characidae most generalized; other families have specialized skeletal structures, jaws, and teeth.

Family Characidae (characins). Fresh to brackish waters; Africa, South and Central America. Tremendous morphological and ecological diversity. Many brilliantly coloured. Variable food habits. Popular aquarium and food fishes. Size 2.5–150 centimeters (1–60 inches). Examples: tetras, piranhas.

Families Erythrinidae, Ctenoluciidae, and Cyndontidae. South America. Large mouths, canine teeth. Erythrinidae lacks adipose fin; Ctenoluciidae has ciliated scales; Cynodontidae, long anal fin. Carnivorous. Food fishes. Size to 120 cm (4 ft).

Family Hepsetidae. Africa. Pikelike; large canine teeth. Carnivorous. Food fishes. Size to 100 cm (40 in.), 55 kilograms (120 pounds).

Family Lebiasinidae. South and Central America. Lateral line and adipose fin usually absent. Small to moderate-sized predators.

Family Gasteropelecidae (hatchetfishes). South and Central America. Deep, strongly compressed body; pectoral fins with well-developed musculature. Capable of true flight. Insectivorous. Aquarium fishes. Size to 10 cm (4 in.).

Family Anostomidae (headstanders). South America. Elongated snout; small mouth with folded or fleshy lips or sucking disk. Head-standing habits. Herbivorous. Aquarium and food fishes. Size to 40 cm (16 in.). The South American families Prochilodontidae (predorsal spine, rough scales), Curimatidae (toothless jaws), and Chilodontidae (specialized pharyngeal teeth) are similar to the Anostomidae.

Family Hemiodontidae (pencil fishes). South and Central America. Lower jaw toothless. Tail-standing posture. Herbivorous. Aquarium fishes. Size to 20 cm (8 in.). Family Parodontidae is similar.

Family Citharinidae (moonfishes). Africa. Deep-bodied, scales often denticulate (toothed), small mouth and teeth. Herbivorous. Aquarium and food fishes. Size to 90 cm (3 ft). The African families Distichodontidae (upper jaw not or scarcely movable) and Ichthyboridae (slender body, upper jaw freely movable, carnivorous) are similar but have ctenoid (ciliate) scales.

Suborder Gymnotoidei

No fossil record. Body elongated; anal fin very long; electric organs present.

Families Gymnotidae (gymnotid eels), *Apteronotidae, and Rhamphichthyidae* (knife fishes). South and Central America. Body greatly compressed, scaled Weak electrical powers. Rhamphichthyidae with elephant-like snout; herbivorous. Other families carnivorous. Size to 90 cm (3 ft).

Family Electrophoridae (electric eels). South America. Body eellike, scaleless. Powerful electric organs. Size to 275 cm (about 9 ft), weight to 22 kg (48 lb).

Suborder Cyprinoidei

Paleocene to present. Mouth toothless, protractile. Adipose fin rarely present.

Family Cyprinidae (minnows and carps). Most in fresh but some in brackish water; Asia, Europe, Africa, North America. Pharyngeal teeth in 1 to 3 rows. Some with 1 or 2 pairs of small barbels. Food habits variable. Food fishes of sport and commercial value; aquarium fishes. Size 2.5–250 cm (1 in. to more than 8 ft). Examples: minnows, carp, goldfish, barb, bitterling.

Family Catostomidae (suckers). North America, Asia. Protractile, sucking mouth on underside of head. Detritus feeders. Food fishes. Size to 90 cm (36 in.).

Families Gyrinocheilidae (algae eaters), *Psilorhynchidae, and Homalopteridae* (hill-stream loaches). Mountain streams, Asia. Adaptations to fast currents include fleshy, suctorial mouth and inhalant–exhalant gill openings (Gyrinocheilidae); ventral sucking disk formed by paired fins (Homalopteridae). Algae feeders. Size to 10 cm (4 in.).

Family Cobitidae (loaches). Asia, Europe, Africa. Wormlike;

scales minute or absent; barbels 3–6 pairs. Intestine sometimes modified for aerial respiration. Mostly carnivorous. Aquarium fishes. Size to 30 cm (12 in.).

Order Siluriformes (catfishes)

Paleocene to present. Body naked or covered with bony plates; adipose fin usually present; pectoral and dorsal fins often with spines. Mostly omnivorous. About 2,500 species.

Family Diplomystidae. South America. One pair of barbels; primitive Weberian apparatus. Size to 24 cm (9½ in.).

Family Ictaluridae (North American freshwater catfishes). Few enter brackish water. North America; widely introduced. Barbels 4 pairs; some with venom glands. Valuable food fishes (sport and commercial). Size to 170 cm (67 in.), 50 kg (110 lb). Examples: bullhead, channel catfish.

Family Bagridae. Found in Asia and Africa. Similar to Ictaluridae but with elongated adipose fin. Food, aquarium fishes. Size to 90 cm.

Family Siluridae. Asia, Europe, Africa. Body compressed; adipose fin lacking, anal fin very long; short dorsal fin (often lacking) without spine. Food; aquarium fishes. Size to 400 cm (13 ft), 300 kg (660 lb). Examples: wels, glass catfish.

Family Schilbeidae. Asia and Africa. Similar to Siluridae, but with adipose fin usually present and spine in dorsal fin. Food fishes. Size to 230 cm (91 in.), 110 kg (240 lb).

Families Amblycipitidae and Akysidae. Asia. Similar to Bagridae but with reduced gas bladder. Akysids inhabit mountain streams, have tuberculated skin. Small size.

Family Amphiliidae. Africa. Similar to Bagridae, but paired fins expanding horizontally for adhesion in fast currents. Size to 21 cm (8½ in.).

Family Sisoridae (mountain-stream catfishes). Asia. Ventral surface flat; thorax with longitudinal plates or adhesive organ. Size to 30 cm (12 in.).

Family Clariidae (labyrinthic catfishes). Asia, Africa; widely introduced elsewhere. Long dorsal and anal fins without spines; adipose fin usually lacking. Treelike air-breathing organ. Food fishes. Size to 130 cm (51 in.). Example: walking catfish. The similar family Heteropneustidae has long, hollow air sacs.

Families Cranoglanididae, Pangasiidae, Chacidae, and Olyridae. Small Asian families, each containing 1 to several species.

Family Malapteruridae (electric catfishes). Africa. Rayed dorsal fin lacking; spines lacking. Electric organs. Food fishes. Size to 120 cm (47 in.), 23 kg (50 lb).

Family Mochokidae (upside-down catfishes). Africa. Bony shield on head and nape. Some swim upside down. Food fishes. Size to 60 cm (24 in.).

Families Ariidae and Plotosidae (sea catfishes). Marine, a few entering freshwater. Tropical coasts; Plotosidae restricted to Indo-Pacific. Nasal barbels lacking; oral incubation of eggs (Ariidae). Adipose fin lacking; long anal and caudal fins confluent (Plotosidae). Food fishes. Size to 115 cm (45 in.).

Family Doradidae (thorny catfishes). South America. Overlapping plates cover sides of body. Intestinal modifications for aerial respiration. Aquarium fishes. Generally small, to 100 cm (40 in.). The related family, Auchenipteridae, has naked flanks and apparent internal fertilization.

Family Aspredinidae (banjo catfishes). A few enter brackish waters and salt waters. South America. Adipose lacking; broad, flat head; large tubercles on naked body. Aquarium fishes. Size to 30 cm (12 in.).

Family Pimelodidae. South and Central America. Similar to Bagridae but lack nasal barbels. Food, aquarium fishes. Size to 130 cm (51 in.), 65 kg (145 lb). Families Ageneiosidae (maxillary barbels only), Hypophthalmidae (1 species, toothless), and Helogeneidae (1 species, no dorsal spine) are South American families similar to the Pimelodidae.

Family Trichomycteridae (parasitic catfishes). South America. Operculum (gill cover) usually with spines. Many parasitic. Size to 10 cm (4 in.). Example: candiru. The similar family Cetopsidae lacks opercular spines.

Families Callichthyidae (mailed catfishes), *Loricariidae* (armoured catfishes), *and Astroblepidae.* South and Central America. Two longitudinal series of overlapping bony plates in Callichthyidae. Three or 4 rows of bony scutes in Loricariidae. Skin naked in Astroblepidae. Sucking mouth (Loricariidae), or mouth and fins modified for adhesion to rocks in mountain streams (Astroblepidae). All herbivores, closely related. Aquarium fishes. Size to 75 cm (30 in.).

Critical appraisal. Ostariophysians are relatively primitive bony fishes, singularly distinct from all other fishes except the Gonorynchiformes. Their specialized Weberian apparatus precludes their having given rise to any higher groups. Differences among various classifications of the superorder are not as great as they appear; the same or similar subgroups are widely recognized, but they may be assigned to different levels in the taxonomic hierarchy. L.S. Berg placed all ostariophysians in one order (Cypriniformes) with two divisions (Cyprini and Siluri). The many families of characins listed by Greenwood are recognized as subfamilies under the single family Characidae by many authorities. Present consensus favours the characoids as the most primitive suborder, an ancestral stock giving rise to cyprinoids in southeastern Asia and to gymnotoids in South America.

The siluriform fishes are more highly specialized than the cypriniforms, and the diplomystids undoubtedly are the most primitive of the catfishes. There is little agreement on the relationships of the other families, but recent research on the caudal skeleton indicates that the Ictaluridae, Bagridae, and Schilbeidae (among others) tend to retain primitive characteristics. Advanced features indicate a relationship between the Clariidae and Heteropneustidae; the Doradidae and Auchenipteridae; the Loricariidae, Astroblepidae, and Callichthyidae; and the Plotosidae and Chacidae. Extensive studies of morphological and other genetic characters of both cypriniform and siluriform fishes are needed before a satisfactory classification and phylogeny can be achieved. (R.W.Y.)

Toadfishes, trout-perches, codfishes, and allies (Paracanthopterygii)

The fishes that constitute the superorder Paracanthopterygii are a predatory, primarily marine group that forms one of about six major branches of the Teleostei, or bony fishes, the dominant modern aquatic vertebrates. Approximately 1,160 living species of paracanthopterygian fishes have been described; they range in length from just a few centimetres to roughly two metres (more than six feet).

In general body form there is considerable diversity, but ichthyologists have classed the Paracanthopterygii as a discrete group, largely on the basis of a distinctive musculature of the jaws, the structure of the caudal (*i.e.,* at the tail end) vertebrae, and the placement of the pelvic fins (they are usually in the midbody region or even farther toward the head).

The Paracanthopterygii comprises six orders: Batrachoidiformes, or toadfishes, about 45 species; Gadiformes, or codfishes, about 800 species; Gobiesociformes, or clingfishes, about 100 species; Lophiiformes, or anglerfishes, about 210 species; Percopsiformes, or trout-perches, about eight species; and Polymixiiformes, or beardfishes, three species. Most of the orders are primarily marine, with worldwide distribution; the percopsiforms, however, occur only in fresh waters of North America. Batrachoidiforms and gobiesociforms occur mainly in tropical and temperate shallow water along continental coasts and to a limited extent in fresh water. Gadiforms are represented by both shallow-water and deep-sea types. The most widely known gadiforms are the commercially important species and the only economically important paracanthopterygians: the true cods (*Gadus*), hakes (*Merluccius, Urophycis*), haddocks (*Melanogrammus*), pollocks (*Pollachius*), and whitings (*Merlangius*). All are abundant in waters of the continental shelf of the North Atlantic, where they have been commercially fished for centuries from both Europe and North America. Lophiiforms live in shallow waters of tropical reefs as well as in the ocean depths. Polymixiiforms occur at moderate depths in most warm seas, generally near continents.

Size range

The largest of the Paracanthopterygii are the codfishes, which grow to about two metres in length and attain weights that may exceed 90 kilograms (about 200 pounds). Certain goosefishes (Lophiiformes) reach a length of about two metres and a body weight of 35 kilograms (about 75 pounds); other lophiiforms are as small as 2½ centimetres (about one inch) long. Batrachoidiforms grow to about 30 centimetres (one foot) in length, gobiesociforms to about eight centimetres (three inches). The largest percopsiforms are about 15 centimetres (six inches) long. Polymixiiforms reach no more than 30 centimetres in length.

NATURAL HISTORY

Life cycle and reproduction. Eggs of the oyster toadfish (*Opsanus tau*) of the western Atlantic—one of the most carefully studied batrachiforms—are laid in dark recesses of all sorts, including sunken tin cans and shoes. The male guards the eggs and young for about three weeks, after which the young fishes begin life on their own. The fish gets its name from the fact that some have been found living in living oysters. Luminous organs known as photophores, numbering several hundred and set in long horizontal rows, are believed to be sexual attractants in the midshipman (*Porichthys*)—so named because the organs resemble rows of bright buttons on a naval uniform. The northern midshipman (*P. notatus*), a common species on the eastern Pacific coast, spawns in shallow water, attaching its eggs to a rocky surface. The male guards the eggs. Like other batrachoidiforms, the midshipman lives and grows on the ocean bottom.

Parasitic habit of pearlfish

Most species of codfishes (which comprise some 70 species of Gadiformes) migrate over long distances. They gather in late winter and early spring to spawn, each species going to a particular area. The periodic movements are closely related to seasonal variations in water temperature. Fecundity of some codfish species is prodigious. The European ling (*Molva molva*) may deposit as many as 60,000,000 eggs each season. The eggs and larvae of most species are found in the plankton (*i.e.*, the aquatic organisms, collectively, suspended in the sea). Weeks or months elapse before the eggs hatch. Young codfishes are commonly found in very shallow water, but they move into deeper water as they become older. The eggs of grenadiers (family Macrouridae), a bottom-feeding group of cods, are believed to be laid near the bottom; the buoyant eggs rise part way to the surface. The larvae are known mainly from below 100 fathoms (about 180 metres, or 600 feet); older larvae occur at greater depths. In the Mediterranean pearlfish (*Carapus acus*), a member of another codlike group (family Carapidae), clumps of eggs, released by the female in late summer, appear at the surface and hatch into a specialized larva, the vexillifer, which lives amid the plankton. After attaining a length of about seven to eight centimetres (about three inches), it transforms to another larval stage, the tenuis, descends to the bottom, and becomes a parasite in a sea cucumber (*Holothuria tubulosa* or *Stichopus regalis*). The tenuis, apparently dependent upon its host for survival, undergoes a further transformation to the juvenile stage; in the process, its length decreases from 20 to 10 centimetres (eight to four inches). The Mediterranean pearlfish is believed to pass most of its life in the host. Very little is known of the general biology of reproductive habits of the brotulas and cusk eels (family Ophidiidae), also of the cod group. They are both oviparous (egg-laying) and viviparous (live-bearing). The males of some viviparous species produce spermatophores (sperm cases). The European eelpout (*Zoarces viviparus*) of the cod family Zoarcidae bears living young about five centimetres in length and numbering as many as 400. Fertilization is internal, and embryonic development occurs in the ovary of the female. Other eelpouts are believed to be live-bearing, but the ocean pout (*Macrozoarces americanus*) of the western Atlantic lays eggs that are guarded by one or both parents.

Drawing by J. Helmer based on (cave fish) N.B. Marshall, *The Life of Fishes* (1965), Weidenfeld and Nicolson Co. Ltd., London, (all but *Polymixia*) David Starr Jordan, *A Guide to the Study of Fishes*, Holt, Rinehart & Winston, Inc.

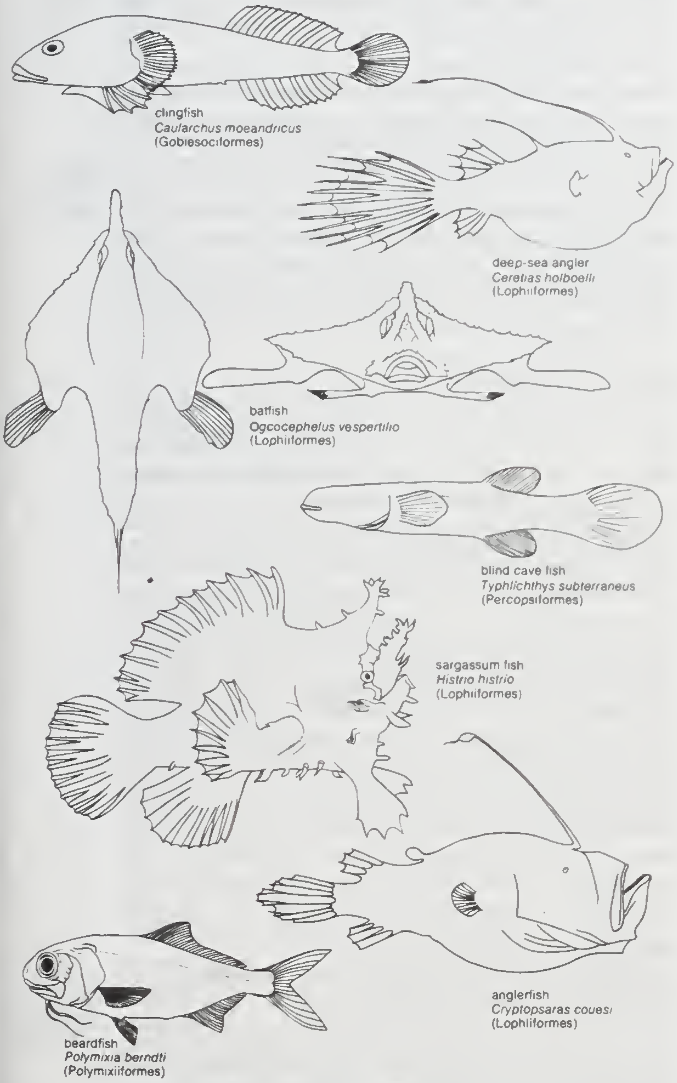


Figure 27: Body plans of representative members of the orders Gobiesociformes, Lophiiformes, Percopsiformes, and Polymixiiformes.

Parasitism of male anglerfish

The lophiiforms are primarily bottom fishes as adults, but many produce floating rafts of eggs. The eggs of the deep-sea anglerfishes (suborder Ceratioidei) are unknown; but it is believed that they float to the surface; the larvae occur in surface waters, gradually descending to deeper waters as they grow older. The females of the deep-sea anglers are from three to 13 times as large as the males. Females have an illicium, or "fishing pole," which is a modified spine of the dorsal, or back, fin that has moved forward onto the top of the head. At the tip of the illicium is a fleshy enlargement, the esca, used to lure prey within range of capture. (The illicium and esca are generally present also in male anglerfish other than in the Ceratioidei.) Commonly the esca is luminous; the female also has other light-producing organs. In 1922 a specimen of the anglerfish *Ceratias holboelli* was discovered; small specimens attached to its abdomen were thought to be its young. A few years later similar finds led to the discovery that the smaller fish were really mature males living parasitically on the female. Further investigation showed that the males, soon after their transformation from the larval state, bite onto an older, larger female, after which the female and male tissues unite; the separate circulatory systems join; and the male becomes a permanent appendage of the female.

Little is known of the reproductive habits of the gobiesociforms, percopsiforms, or polymixiiforms. Gobiesociforms are known to lay eggs in shallow water, attaching them to rocks or plants; and percopsiforms are known to spawn in the spring of the year in shallow water.

Ecology and behaviour. All batrachoidiforms are bottom dwellers. True toadfishes (Batrachoidinae, about 25 species) occur in shallow or moderate depths along continental coasts; some ascend rivers. The oyster toadfish lives under rocks or amid debris, awaiting prey of almost any type, which is taken with a sudden snap. Venomous toadfishes (Thalassophryninae, about nine species) are restricted to the coasts and rivers of Central and South America. Because of their sluggish habits, these fishes are sometimes stepped on by man and can inflict painful wounds. Midshipmen (Porichthyinae, about 12 species), which are restricted to tropical and temperate coasts of

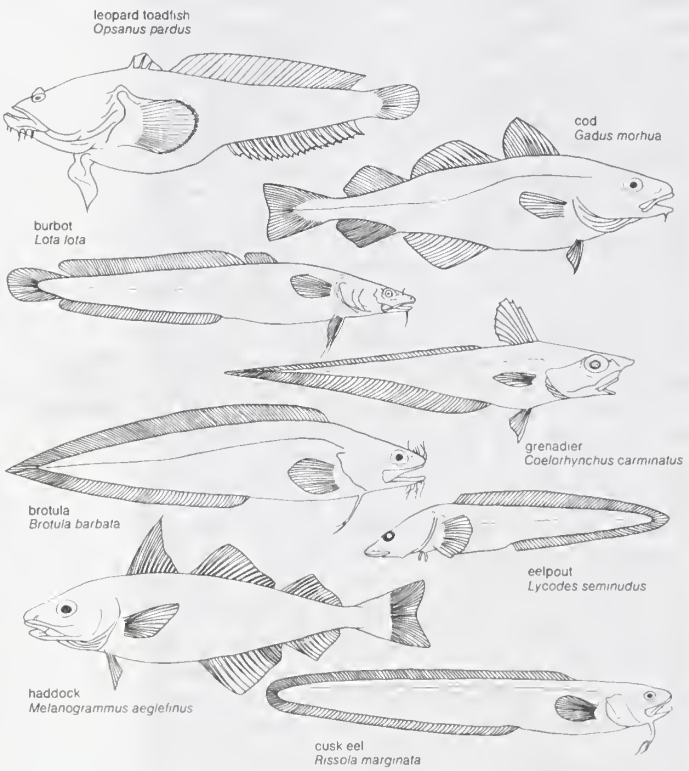


Figure 28: Body plans of representative members of the orders Batrachoidiformes (toadfish) and Gadiformes.

Drawing by J. Helmer from David Starr Jordan, *A Guide to the Study of Fishes*, Holt Rinehart & Winston, Inc

the Americas, are unusual in being shallow-water fishes with photophores, a feature generally found in deepwater forms. Most midshipmen occur in depths of less than 50 fathoms (one fathom = six feet), and all are found in water shallower than 200 fathoms.

Gadiform fishes of the family Gadidae (about 70 species) are all marine species, except for the burbot (*Lota lota*); some, however, ascend rivers with the tides. Bottom dwellers, they occur on the continental shelves from shallow water to about 200 fathoms and, although distributed throughout the oceans, are most numerous in the eastern North Atlantic. Deep-sea cods (Moridae, about 70 species) are cold-water bottom fishes, living at greater depths along the continental slopes. Grenadiers (Macrouridae, about 300 species), typically bottom fishes, live along the continental slopes at depths of 100 to 1,000 fathoms. Few species are cosmopolitan in distribution, but the group as a whole is widely distributed in tropical and temperate latitudes. A few species of gadiforms (Muraenolepididae, four species) are confined to Antarctic seas, and, like the cods, they are bottom fishes, living at moderate depths. The pearlfishes (Carapidae, about 27 species) are marine, mainly tropical, shallow-water, eellike fishes adapted to living inside the body of various invertebrates. They have been collected from a variety of hosts, including tunicates, oysters, and sea cucumbers. Brotulas and cusk eels (Ophidiidae, about 250 species) are mainly bottom dwellers. Some are shallow-water species with nocturnal habits, but the group as a whole is the dominant teleostean family at depths greater than 2,000 fathoms. Some have been taken at depths of about 4,000 fathoms, the greatest depth at which any form of fish life is known. Eelpouts (Zoarcidae, about 80 species) are bottom fishes, commonly occurring from shallow water to depths of 1,000 fathoms. Species are most abundant in the higher latitudes of both hemispheres, especially in the north. Eelpouts are common in shallow water of Arctic and Antarctic seas.

Eelpouts

Gobiesociforms (about 100 species) are mostly marine fishes, typically inhabiting the intertidal zone. Some species (*Diademichthys*) hide among the spines of sea urchins. In tropical America, four species (*Gobiesox*) are known from swift-flowing freshwater streams.

Of the lophiiforms, the ceratioids, or deep-sea anglerfishes, are the only abyssal (deep-sea) forms. They occur primarily at depths of 1,000 to 3,000 fathoms. Unlike other lophiiforms, they are midwater fishes, are uniformly black, and have no pelvic fins. They are apparently feeble swimmers, depending primarily on their light organs to attract prey. Some (*Melanocetus, Linophryne*) are known to swallow fishes several times their own length, accommodating them in a highly distensible stomach. Crustaceans and other invertebrates are also eaten. Many lophiiforms—so-called·frogfishes (Antennariidae, about 60 species)—are shallow-water forms, commonly inhabiting coral reefs. Frogfishes typically have highly varied colour patterns, and some species are able to change colours. In habit they are sedentary but can use their fins to walk on the bottom and to climb over obstacles. The tropical sargassum fish (*Histrio histrio*), so called because it lives amid floating brown algae of the genus *Sargassum*, clings to the branches of algae with prehensile (*i.e.*, adapted for seizing, or wrapping around) pectoral fins as it searches for prey, which is sucked into the mouth by the powerful jaws and expandable cheeks.

The lophiiform group known as goosefishes (Lophiidae, about 12 species) seldom occur in shallow water, preferring instead the moderate depths (10 to 500 fathoms) along the continental slopes of tropical and temperate region. The batfishes (Ogcocephalidae, about 60 species), are mainly deepwater lophiiforms, but some (*Ogcocephalus, Halieutichthys*) are regularly found in water only a few feet deep. Like frogfishes, they walk on the bottom, using their pectoral fins. Batfishes are awkward swimmers and, when disturbed, tend to bury themselves in the bottom rather than swim away.

Percopsiforms (about eight species) live under conditions of dim light. Cave fishes, with eyes reduced to nonfunctional rudiments, have elaborate systems of sense organs in the skin of the head, body, and tail; they live in total darkness. Because of the secretive habits of percopsiforms, little is known about species other than the trout-perch (*Percopsis omiscomaycus*), which is widely distributed in central North America and is abundant in some of the Great Lakes, where it occurs in clear water to a depth of about 35 fathoms. Polymixiiforms, numbering only three marine species, are generally found at depths of 150 to 350 fathoms.

Fishes without eyes

FORM AND FUNCTION

Batrachoidiforms generally have two dorsal fins; a small anterior fin, usually with two spines; and a long posterior fin. In venomous species, the hollow fin spines form an efficient apparatus for the injection of venom. A similar spine is found on each cheek (operculum).

Among the gadiforms, the dorsal and anal fins of some deep-sea cods are distinctively arranged as three dorsals and two anals. This arrangement also occurs in some codfishes (Gadidae). The macrourids are characterized by a long, tapering tail. A tubular light organ containing luminescent bacteria is sometimes present along the ventral midline of both sexes. All but a few species of macrourids have a well-developed swim bladder; in the males of some species and in some codfishes, the swim bladder is equipped with drumming muscles, indicating that sound can be produced. In the bregmacerotids and muraenolepidids there are two dorsal fins, with the anterior fin represented by a single ray. Brotulas may have the pelvic fins either present or absent, but cusk eels have them anterior in position, under the lower jaw. They are kept in continuous probing motion, as the fish swims just off the bottom, and aid in detecting food. Zoarcids (eelpouts) are elongated, eel-like fishes. Their pelvic fins are either rudimentary or entirely absent.

Gobiesociforms, with a depressed head, wide mouth, and tapering body, resemble toadfishes, but they are distinctive in having a prominent sucking disk on the ventral surface. The paired pelvic fins, thoracic in position, form part of the disk, various fleshy pads and folds forming the remainder. The disk allows clingfishes to hold fast to rocky bottoms amid the often turbulent wave action of their shallow-water environment.

Suction disk for attachment to bottom

Some lophiiforms are unique among teleostean fishes in having only two gills. The ogcocephalids are somewhat flattened anglers, in this respect resembling lophiids rather than the ballon-like antennariids; they are distinctive in having the illicium, when not in use, concealed in a tube (illicial cavity) between the eyes and over the mouth. Like most anglerfishes they lack typical scales but are distinctively equipped with bony tubercles (projections) and spines imbedded in the skin.

The polymixiiforms are singular in having a pair of fleshy barbels, or "whiskers," under the jaw. Each barbel is supported by three small bones.

EVOLUTION AND PALEONTOLOGY

Fossil batrachoidiforms include only material from lower Pliocene marine deposits (about 5,000,000–7,000,000 years old) of North Africa. These fossils are similar to a living species, *Batrachoides didactylus.*

Fossil gadiforms are relatively numerous and are known primarily from Tertiary marine deposits (about 2,500,000–65,000,000 years old) of the Northern Hemisphere. A Paleocene fossil (54,000,000–65,000,000 years old) has been identified as a codlike fish; some Eocene fossils (38,000,000–54,000,000 years old) have been identified for the families Bregmacerotidae and Gadidae; and Oligocene–Miocene fossils (7,000,000–38,000,000 years old) for the families Bregmacerotidae, Gadidae, Macrouridae, and Ophidiidae. In addition, many fossil ear stones (otoliths) and scales, beginning with specimens from the Cretaceous (65,000,000–136,000,000 years ago), are similar to the Gadiformes. Fossil gobiesociforms are unknown. Fossil lophiiforms include two species from Eocene marine deposits of Europe and one species from Pliocene marine deposits of North Africa; one Eocene species has been identified as a goosefish (Lophiidae), the other as a frogfish (Antennariidae).

Fossil percopsiforms include three genera from Tertiary freshwater deposits of North America and one (*Sphenocephalus*) from Cretaceous marine deposits of Europe. Of the North American genera, two (*Amphiplaga, Erismatopterus* from the middle Eocene) have been identified as trout-perches (Percopsidae), and one (*Tricophanes,* Oligocene–Miocene) as a pirate perch (Aphredoderidae). The relationships of *Sphenocephalus* are obscure. Fossil polymixiiforms include a diversified group of about six genera known primarily from Cretaceous marine deposits of Europe and the Middle East; a few others are known from the Tertiary.

CLASSIFICATION

Annotated classification.

SUPERORDER PARACANTHOPTERYGII

Most with a distinctive type of jaw musculature (involving levator maxillae superioris muscle and associated structures); pelvic fins usually placed anteriorly, thoracic (midbody) or even further forward; primarily marine; worldwide distribution; about 1,600 living species.

Order Polymixiiformes (beardfishes)

Middle Cretaceous to Recent. Barbels supported by rays; spines on the dorsal and anal fins; pelvic fins subthoracic. Deepwater marine fishes; three species. Adult length about 30 cm (12 in.).

Order Percopsiformes (trout-perches, pirate perches, and cave fishes)

Eocene to Recent. Mouth gape and buccal dentition reduced; median fin spines reduced or lost; head with spine ornamentation. About 8 living species, all freshwater; North America; length 8–15 cm (3 to 6 in.).

Order Gadiformes (cods, cusk eels, pearlfishes, eelpouts, and grenadiers)

Paleocene to Recent. Early gadiforms were similar in structure to early percopsiforms, but almost all remained marine and subsequently specialized into a variety of environments. Reduced caudal skeleton; elongate body; altered head and jaw structure. Very reduced fin spines; marine, worldwide. About 800 species. Length 7 to about 200 cm (2¾ in. to 6½ ft).

Order Batrachoidiformes (toadfishes)

Miocene to Recent. Bottom fishes with short, small, spinous dorsal fins; long soft-rayed dorsal fins; flat heads; about 45 species; marine, occasionally freshwater, shore fishes of tropics. Length to about 30 cm (12 in.).

Order Lophiiformes (goosefishes, anglerfishes, frogfishes, and batfishes)

Eocene to Recent. Spinous dorsal fin modified as a movable lure. Some deep-sea forms with light organs and males parasitic on females. Marine, widespread; in shallow-water and deep-sea habitats. About 210 species. Length to about 200 cm (6½ ft).

Order Gobiesociformes (clingfishes)

Recent; flattened, depressed fishes with a ventral sucker formed of the pelvic fins and surrounding tissue; no spiny dorsal fin; about 100 species; marine and occasionally freshwater in tropics and along many temperate seacoasts.

Critical appraisal. The interrelationships of the groups listed here as paracanthopterygians are not yet well established, and the classification given here is provisional. There is considerable agreement that trout-perches (Percopsiformes) and cods (suborder Gadoidei) are closely related, and this agreement may be considered the basis of the group Paracanthopterygii. What other fishes should be included in the Paracanthopterygii is a question receiving continued study. Some ichthyologists have held that clingfishes (Gobiesociformes) are related to dragonets (Callionymidae); that eelpouts (Zoarcidae) are related to blennies (Bathymasteridae, Blenniidae, etc.); that brotulas, cusk eels, and pearlfishes (suborder Ophidioidei) are related to the river blackfish of Australia (the perchlike *Gadopsis*); that beardfishes (Polymixiiformes) are related to squirrelfishes and their relatives (Beryciformes); and that toadfishes (Batrachoidiformes) and anglers (Lophiiformes) also have their relationships within the great mass of perchlike fishes (Acanthopterygii). In addition, killifishes (Cyprinodontidae) and related live-bearers (Poeciliidae) are believed by some writers to be related to trout-perches (Percopsiformes) and by others to silversides, flying fishes, and their relatives (Atheriniformes). Thus, future study may result in the transfer of some groups from the Paracanthopterygii to the Acanthopterygii and vice versa.

(G.J.N.)

Flying fishes, needlefishes, cyprinodonts, and allies (Atheriniformes)

The order Atheriniformes contains 15 families of marine and freshwater spiny-finned fishes, including the flying fishes, needlefishes, silversides, and cyprinodonts. The last group, the Cyprinodontidae, is an abundant tropical and subtropical family that includes the guppies, mollies, swordtails, and many other aquarium fishes. In addition to the Atheriniformes, this section will treat the three smaller related orders Beryciformes, Zeiformes, and Lampridiformes, the most primitive groups of the superorder Acanthopterygii or spiny-finned fishes.

GENERAL FEATURES

Beryciforms and zeiforms are mostly deep-bodied fishes of small to moderate size, a foot or less in length. The lampridiforms include a few rare, deep-bodied forms, notably the disk-shaped opah, which may reach more than 136 kilograms (300 pounds) in weight, but the majority are much elongated, ribbonlike fishes, including the giant oarfish, *Regalecus,* which reaches eight metres (25 feet) in length and is the probable source of many sea-serpent legends. The atheriniform silversides, flying fishes, needlefishes, and halfbeaks tend to be slender, elongate fishes, up to 0.3 to 0.9 metre (two to three feet) in length. The cyprinodonts and their relatives are diminutive and include some of the smallest vertebrates. Many cyprinodonts are important as experimental animals in biological research and as useful predators in the control of insect-borne diseases.

NATURAL HISTORY

Most beryciforms, zeiforms, and lampridiforms are inhabitants of the open oceans, usually living at considerable depth, and little is known of their natural history. All appear to produce numerous small eggs. The best known of the beryciform groups are the squirrelfishes and soldierfishes (family Holocentridae), abundant around coral reefs in warm seas. Typical of beryciforms, they are red in colour, with large eyes. Holocentrids are nocturnal,

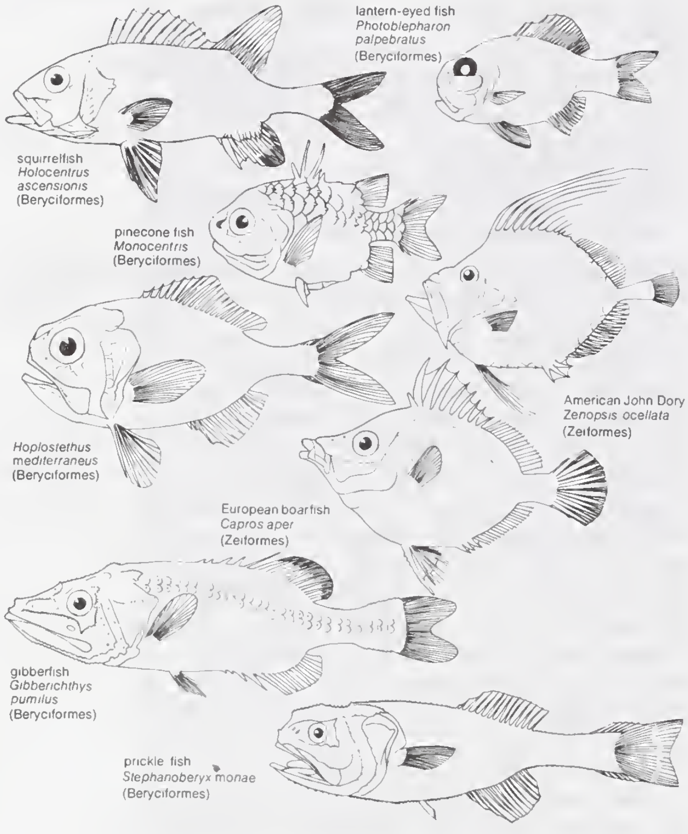


Figure 29: Body plans of Beryciformes and Zeiformes.

From (*Hoplostethus, Stephanoberyx*) *The Fishes of North and Middle America* by David Starr Jordan and Barton Warren Evermann, Bulletin of the U S National Museum No 47, 1900, reprinted by permission of the Smithsonian Institution, (*Gibberichthys*) A E Parr, *Bingham Oceanographic Collections*, vol 14, no 6, (*Photoblepharon, Monocentris*) P P Grasse, *Traite de Zoologie*, vol 13 (1958), Masson et Cie, Editeurs, (*Capros, Holocentrus, Zenopsis*) N B Marshall, *The Life of Fishes* (1965), Weidenfeld & Nicolson, Ltd

Luminescent organs

sheltering in crevices during the day and emerging at night to feed. They are notable sound producers, having special drumming muscles attached to the swim bladder, and many have connections between the swim bladder and the ear to improve hearing: presumably these sounds and their reception play some part in courtship. In holocentrids the young (larva) is quite unlike the adult, with a projecting spiny snout and enlarged spines in front of the gill cover. There is a pronounced metamorphosis (a major change in body plan on reaching maturity). It is probable that some of the deep-sea beryciforms undergo similar metamorphoses; the larva of the fanged *Caulolepis* was for many years placed in a different family from the adult, and the genus *Kasidoron*, recently discovered and placed in a distinct family, may be only the larva of *Gibberichthys*. Another family of beryciforms found near the surface is the Anomalopidae, or "strange-eyes," so-called because of a large luminous organ lying directly below the eye, which is switched on by muscular eversion, turning the inside outward (in *Anomalops*) or by the withdrawal of a pigmented cover (in *Photoblepharon*). The Monocentridae, the bizarre pinecone fishes, are another beryciform group with luminous organs, in this case located on the chin. The majority of beryciforms are generalized predators, but a few coral-reef forms are grazers.

Among zeiforms, at least one species, *Zeus faber*, produces sounds by drumming muscles and breeds inshore. Little is known of the biology of the oceanic forms, but some certainly undergo metamorphosis, especially the Oreosomatidae, whose larvae are studded with large, spinous tubercles. All zeiforms are highly compressed fishes, with stiff bodies and long dorsal and anal fins: probably they swim by undulating these fins rather than by flexing the body. This slow, stealthy mode of swimming, coupled with their highly protrusile mouths, adapts them for stalking and engulfing prey.

The lampridiforms are all oceanic fishes. Metamorphosis is recorded in dealfishes and oarfishes, the young of which have rather deep bodies and greatly elongated finrays. All are slow swimmers, and the larger forms, the opah and the oarfishes, which are characteristic of surface waters, use their protrusile, toothless mouths as traps for small, planktonic (free-floating) organisms. The deep-sea forms have feebly toothed jaws and are predators. A remarkable modification in one lampridiform, *Lophotes*, is the presence of an ink sac, discharging a viscous, black secretion into the hindgut, thence into the water. These fishes probably use their ink as a defense mechanism, as do squids. *Stylephorus*, a highly modified deep-sea lampridiform, has projecting, telescopic eyes.

Flying fishes

Among atheriniforms there is an extraordinary variety of locomotor, reproductive, and ecological adaptations. Locomotor modifications are most marked in the flying fishes, but the origin of the "flying" habit can be traced in flying fish relatives such as the halfbeaks, garfishes, and skippers. All are surface fishes of the open ocean and are capable of leaping or skipping on the surface, sometimes for considerable distances, thus allowing them to escape predators. The tail (caudal) fin is usually asymmetrical, with the lower lobe longer than the upper, and while the body is out of the water the lower lobe vibrates as a scull driving the fish along. True flying fishes have a similar asymmetrical tail, but the pectoral fins are inserted high on the shoulders and are greatly enlarged, with long, stiff fin rays supporting a web of skin. In the most highly evolved flying fishes, the pelvic fins are also enlarged and winglike. The fish accelerates under water by rapid vibration of the tail and fin, with the paired fins furled. On breaking surface, the pectoral fins are expanded, but the lower lobe of the tail remains in the water, sculling rapidly and accelerating the fish. The pelvic fins are then expanded, lifting the tail out of the water and initiating gliding flight. As airspeed is lost, the fish may fall back into the sea or furl its pelvic fins, dropping the lower lobe of the tail into the water and picking up speed for a further glide. Up to five repeated takeoffs have been observed, producing a total flight time of almost half a minute and covering several hundred yards.

Marine atheriniforms are mostly predators, the predatory habit being most highly developed in the garfishes and needlefishes, with their long, formidably toothed jaws. Freshwater atheriniforms are generally adapted for feeding at the surface, on insect larvae and small crustaceans.

All atheriniforms are characterized by the production of few, large, adhesive eggs, by mating in pairs, usually accompanied by sexual dimorphism (*i.e.*, the sexes markedly different), and many groups exhibit various reproductive specializations, the most advanced of which is viviparity (the production of functional young, instead of eggs). The young are normally miniatures of the adult and there is no metamorphosis. Sauries, needlefishes, flying fishes, and marine halfbeaks are pelagic (*i.e.*, inhabiting open ocean) and breed either in the open sea (sauries, flying fishes) or near the shore (needlefishes, halfbeaks), the eggs often attaching to floating objects by adhesive filaments. The freshwater halfbeaks are mostly viviparous and have an elaborate courtship behaviour.

Atheriniforms of the suborder Atherinoidei fall into two groups, the silversides (Atherinidae and their close relatives) and the more specialized phallostethoids. The silversides are mainly freshwater fishes and show some reproductive specializations in courtship behaviour and sexual dimorphism (coloration and fin shape). They breed near the shore, attaching the eggs to plants. The grunion (*Leuresthes tenuis*) breeds on the California coast, schooling in the surf at extreme spring high water and spawning on the shore, where the female buries the eggs in the sand. The eggs hatch when they are exposed by the next spring tide, two weeks later. In phallostethoids, males have a fleshy, asymmetrical intromittent organ, the priapium, under the throat, formed from the modified pelvic fins. Although fertilization is internal, viviparity is not known to occur.

Breeding specializations of cyprinodonts

Even the most primitive atheriniforms in the suborder Cyprinodontoidei show the usual reproductive specializations of the group: sexual dimorphism and complex behaviour patterns in courtship and spawning. In the Mexican topminnows (Goodeidae) viviparity has developed, the embryos absorbing nourishment within the oviduct of the mother by means of threadlike outgrowths. In the livebearers (Poeciliidae), an abundant group in the American tropics and subtropics, sexual dimorphism affects many parts of the body. Males have a complex intromittent organ, the gonopodium, formed of modified anal fin rays. One member of the group is oviparous, shedding the eggs while the embryo is only partially developed, but in the guppies, mollies, and swordtails, where the male is much smaller than the female and more brightly coloured, the young are born fully developed, and a series of broods, at about monthly intervals, may result from a single fertilization. In wild cyprinodont populations the sex ratio is frequently unusual, with many females to each male. In *Jenynsia* and *Anableps* (the four-eyed fish) the gonopodium and female reproductive opening are asymmetrical. Both dextral and sinistral forms occur within a species, dextral males mating with sinistral females and vice versa.

From (*Lampris regius*) N.B. Marshall, *The Life of Fishes* (1965), Weidenfeld & Nicolson, (*Mirapinna esau*) P.P. Grasse, *Traite de Zoologie*, Masson et Cie, Paris, (*Kirtlandia vagrans*) D.S. Jordan, *The Study of Fishes*, (others) *The Fishes of North and Middle America* by David Starr Jordan and Barton Warren Evermann, Bulletin of the U.S. National Museum No. 47 (1900), reprinted by permission of the Smithsonian Institution

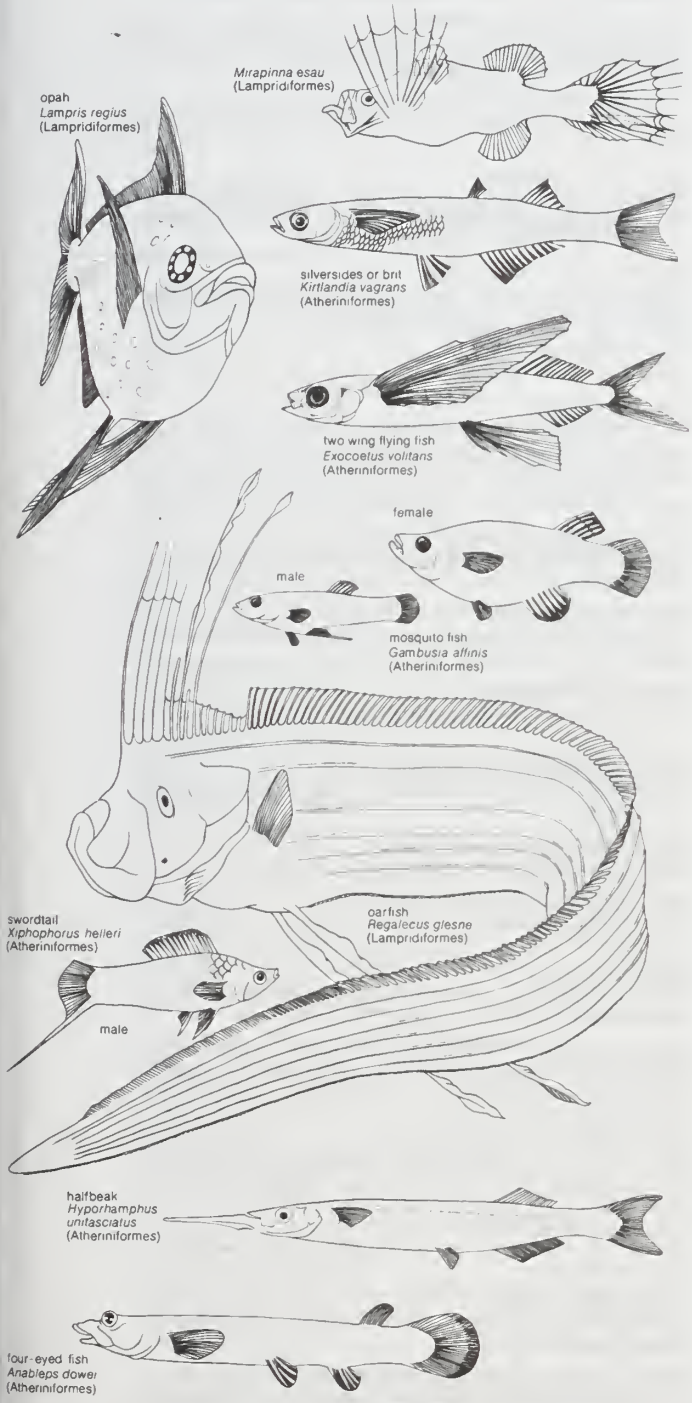


Figure 30: Body plans of Lampridiformes and Atheriniformes.

Ecological adaptations in atheriniforms are most marked in freshwater species. Cyprinodonts are among the hardiest of fishes and survive in the most rigorous environments. Some cyprinodonts have become adapted to life in hot springs in Africa and America and seem capable of surviving water temperatures approaching the coagulation point of protoplasm. Others survive in stagnant, almost or completely deoxygenated waters, either by taking in water at the surface film, or by breaking surface and gulping air, although no accessory respiratory structures are developed. Some cyprinodonts have overcome the rigours of a seasonal tropical habitat by becoming annuals, growing rapidly and reaching sexual maturity in small temporary bodies of water during the wet season, and on the approach of the dry season, mating and burying the eggs in the mud. The eggs can survive droughts for up to five years, hatching rapidly with the onset of the succeeding wet season. Perhaps another response to rigorous environments is the occurrence in some cyprinodont populations of functional hermaphrodites, capable of self-fertilization and hence of maintaining a population from one surviving parent.

FORM AND FUNCTION

The fishes discussed here share a number of anatomical features typical of the more advanced teleosts. These include a closed swim bladder; separation of the parietal bones by the supraoccipital; jaws that protrude to some extent, with the maxillary bone (toothless except in a few beryciforms) acting as a lever to move the large premaxilla; the pectoral fins inserted high on the flank and the pectoral girdle without a mesocoracoid arch; and a tail skeleton supported by two or less vertebrae. Otherwise, there is considerable structural variation.

Structure of beryciforms

Beryciforms are the most primitive fishes of the four groups under discussion, exhibiting primitive features: the presence of two supramaxillary bones in the upper jaw; an orbitosphenoid bone between the eyes; a tail fin containing 19 principal rays, which insert on six hypural bones supported, in turn, by two vertebrae. Occasionally, they have teeth on the maxillary bone (in the modern holocentrid *Myripristis* and a few Cretaceous fossils). There are many ways in which beryciforms approach the perciforms, the typical "spiny-rayed" fishes. Such resemblances are seen in a number of features: the structure of the mouth, with a normal acanthopterygian pattern of jaw muscles and ligaments; the spiny head bones and ctenoid scales (with a serrated edge); a projection called a subocular shelf on the bones below the eye; stout spines in front of the dorsal, anal, and pelvic fins; bony contact between the pelvic and pectoral girdles; and the short, deep trunk, with about 25 vertebrae. The more generalized beryciforms (holocentrids, trachichthyids, and berycids) exhibit all of these features, but in several lineages degeneration has occurred, associated with life in the deep. In such deep-sea beryciforms as the big-scale fishes (Melamphaeidae), fin spines tend to be absent, the pelvic fins have moved back to the abdomen, and the head bones and scales have become thin and flimsy. Also, in some species primitive structures such as the orbitosphenoid and supramaxillary bones are lacking, and fusions within the tail skeleton have resulted in a condition resembling that of perciforms. The swim bladder is reduced or lost in some.

Structurc of zeiforms

Anatomically, the zeiforms resemble perciforms more closely. Almost the only feature that distinguishes zeiforms from perciforms is the presence, in the former, of two or three more rays in the pelvic fins, and in some zeiforms even this distinction fails to hold. Nevertheless, these extra pelvic rays and a few other features, notably the structure of the otoliths ("ear stones," used in maintaining balance), indicate that the zeiforms are beryciform relatives that have independently attained the perciform evolutionary level. Typically, the zeiform has a highly protrusible mouth, a separate spinous dorsal fin, ctenoid scales, and a short, deep trunk; the most primitive members of the order have 24 or less vertebrae.

The most primitive lampridiforms are also deep-bodied fishes, with spines in front of the dorsal and anal fins, the

Structure of lampridiforms

pelvic fins directly below the pectorals, an orbitosphenoid bone in the skull, and a tail fin with 19 principal rays, in which they resemble beryciforms. Lampridiforms differ from beryciforms, however, in never having a subocular shelf or pelvic spine, in having more numerous vertebrae, and in having the upper tail fin supports fused with an independent vertebral centrum, a condition resembling that found in the cods and their relatives (Paracanthopterygii). Most lampridiforms have highly protrusile jaws in which depression of the lower jaw dislocates the maxilla of the upper, so that it moves forward bodily, carrying the premaxilla with it. This is a different method from that adopted by other acanthopterygians, hence the name allotriognaths ("strange-jaws") originally applied to the group. A parallel can be drawn between the beryciforms and lampridiforms in certain modifications exhibited by the deep-sea forms, compared with their surface-living relatives. These include the loss of fin spines, reduction in ossification, and reduction of the swim bladder. The most striking features of the more highly evolved lampridiforms, however, are peculiar to the group: great elongation of the trunk, accomplished by increase in vertebral number and elongation of the vertebrae themselves, and reduction of the tail to a small, asymmetrical or filamentous appendage.

Structure of atheriniforms

The atheriniforms are an extremely varied group. There are many structural resemblances to more advanced acanthopterygians, but these are in mosaic distribution, indicating that most have been independently acquired. The jaws of many atheriniforms are protrusile, but the structural modifications by which this is achieved are quite different from those of typical acanthopterygians. The simple, shelflike head of the maxillary bone is attached to the palate only by ligaments, not by a mobile joint. The premaxilla is longer than the maxilla and also has a simple head. Protrusion of the jaws is accomplished by twisting the maxilla and displacing its head forward; the complex system of joints and ligaments characteristic of other acanthopterygians is not developed. The palate is usually toothless, and the series of infraorbital bones incomplete, only the first (lachrymal) and last (dermosphenotic) bones being present. The skull bones are not spiny, but the scales are often ctenoid. The pelvic girdle may have a ligamentous connection with the shoulder girdle but often lies further back, and the girdles never acquire the direct contact that characterizes higher acanthopterygians. The pelvic fin has six or fewer rays, but there is no pelvic spine. The atheriniform tail skeleton is of an advanced type, usually with two large plates emanating from a single supporting centrum, as in some advanced perciforms. The caudal fin contains 17 or less principal rays. There are a few spines in front of the dorsal and anal fins in many atheriniforms, and the members of the Atherinidae and Phallostethidae have a small, separate spinous dorsal fin, but atheriniform spines appear to have evolved independently from those of true acanthopterygians.

An extreme example of adaptation to life near the air-water interface, the habitat of most atheriniforms, is the eye of *Anableps,* the four-eyed fish, so named because each eye is a double structure. The eye is set high on the head and the upper part projects above the water. The cornea is divided by a horizontal band of pigment, separating an upper, strongly convex part from a lower, flatter division. The iris has a pair of projections partially dividing the pupil into two, and the upper is effective for aerial vision, the lower for underwater vision.

Much work has been done on the genetics of atheriniforms, perhaps the most surprising result being the hatching of hybrids between *Fundulus* (Cyprinodontidae) and *Menidia* (Atherinidae), fishes placed in separate suborders. A physiological peculiarity of some marine atheriniforms, garfishes and needlefishes, is a bright green coloration of the bones and even the flesh, due to retention of a bile pigment, biliverdin.

EVOLUTION, PALEONTOLOGY, AND CLASSIFICATION

Paleontology. The four orders Beryciformes, Zeiformes, Lampridiformes, and Atheriniformes are primitive groups within the superorder Acanthopterygii. The Beryciformes and Zeiformes apparently form a related group, originating in the Cretaceous, its closest relatives being the Perciformes. The Lampridiformes also originated in the Cretaceous and are of uncertain relationships, being to some extent intermediate between the Acanthopterygii and Paracanthopterygii. The Atheriniformes represent a radiation from near the base of the acanthopterygian stock, but their exact relationships within this group are not known. The present distribution of atheriniforms indicates that the group arose in fresh or brackish waters of the tropical Indo-Pacific region, but little is known of their early fossil history.

Annotated classification.

ORDER BERYCIFORMES

Spiny-rayed fishes with a pelvic spine, an orbitosphenoid, and 19 principal rays in the tail. Of the two main lineages the first contains the Holocentridae, coastal fishes of warm seas. The second is a series of oceanic families centring around the Trachichthyidae. Both groups have fossil records back to the Cretaceous, the 2 lines converging in the Middle Cretaceous.

Family Holocentridae (soldierfishes and squirrelfishes)
Circumtropical, with partly separate spinous dorsal fin. Several extinct genera, Middle Cretaceous onward.

Family Monocentridae (pinecone fishes)
Armoured, very spiny. Teeth on endopterygoid bone. Two genera; Indo-Pacific.

Family Trachichthyidae
Midwater (mesopelagic) or deepwater pelagic fishes, worldwide. Skull bones cavernous, with large mucus cavities. Several extinct genera, Middle Cretaceous onward.

Family Berycidae (alfonsinos)
Upper and midwaters in open ocean; worldwide. Pelvic girdle enlarged and tightly joined with the pectoral.

Family Anoplogasteridae
Deep-sea, adults with large fangs; 1 genus.

Family Diretmidae
Very deep bodied, compressed fishes; 1 genus.

Family Anomalopidae (lantern-eyed fishes)
With subocular luminous organ, found near the surface at night; 2 Indo-Pacific genera, 1 Atlantic.

Family Stephanoberycidae (prickle fishes)
Scales and head spiny, fin spines reduced; bathypelagic, worldwide; 3 genera.

Family Melamphaeidae (big-scale fishes)
Abundant deepwater open ocean fishes, worldwide; softbodied and black. Fossils in the Miocene.

Family Gibberichthyidae
Like Melamphaeidae but with stronger fin spines. Atlantic, 1 or 2 genera.

Family Rondeletiidae (whale fishes)
Head large, no scales, fin spines, or swim bladder; bathypelagic, 1 genus.

Family Cetomimidae (whale fishes)
Mouth huge, spineless fins, bathypelagic, worldwide.

Family Barbourisiidae (whale fishes)
No fin spines, scales reduced to minute spines, red, bathypelagic, 1 genus.

ORDER ZEIFORMES

Like Perciformes but with up to 9 pelvic rays and only 12–13 principal caudal rays.

Family Caproidae (boar fishes)
Most primitive family, 21–23 vertebrae, fossils in the Oligocene; 2 genera, worldwide.

Family Zeidae (John Dories)
Deep bodied and laterally flattened. Mouth large; scales reduced; more than 30 vertebrae. Several genera, worldwide; fossils in the Eocene.

Family Grammicolepididae
Mouth very small, scales drawn out into oblique bands. Two genera, mesopelagic.

Family Oreosomatidae
Larva covered with large tubercles. Four genera, benthic (bottom dwelling); worldwide.

Families Zeniontidae and Macrurocyttidae
Two small families, the first with two genera, the second with one, too poorly known to be characterized.

ORDER LAMPRIDIFORMES

Similar to Beryciformes, but with no pelvic spine; upper hypural bones fused with their supporting centrum.

Suborder Lampridoidei

Deep-bodied forms.

Family Veliferidae

One living genus (*Velifer*) with saillike fins, 33 vertebrae. Fossils from Paleocene and Eocene, several extinct genera.

†*Families Aipichthyidae and Pharmacichthyidae*

Extinct families, each containing a single Upper Cretaceous genus; appear to be primitive lampridiforms, resembling *Velifer* in deep trunk, but with fewer vertebrae and more primitive tail skeletons.

Family Lampridae (opahs)

One genus (*Lampris*); 15–17 pelvic rays, 46 vertebrae. Fossils from Miocene. Length to 2 m (6½ ft), weight to 140 kg (over 300 lb); surface waters (epipelagic) of warm seas; widespread.

Suborder Trachipteroidei

Ribbonlike, about 100 vertebrae.

Family Trachipteridae (dealfishes)

Pelvic fins with 5–9 rays, no anal fin, jaws toothed. Length to 1.2 m (4 ft); epipelagic. Worldwide in warm seas.

Family Lophotidae (unicorn fishes)

Scales lacking; pelvic fins small or absent, anal fin short. Fossils from Oligocene. Worldwide in warm seas.

Family Regalecidae (oarfishes)

Anal fin lacking; 1 pelvic ray elongated; jaws toothless; length to 9 m (30 ft); weight to 300 kg (660 lb). Mesopelagic, tropical.

Suborder Stylephoroidei

Family Stylephoridae

Deep-sea forms with enlarged telescopic eyes, about 50 vertebrae, 2 filamentous caudal rays. Known from only a few specimens.

Suborder Ateleopoidei

Family Ateleopidae

Specialized, deep-sea, bottom-living fishes, Indo-Pacific and Atlantic, usually placed among the primitive teleosts, but probably lampridiform.

Suborder Mirapinnoidei

Families Mirapinnidae and Eutaeniophoridae

Three species of little-known mesopelagic fishes, usually placed as a distinct order of lower teleosts (Mirapinniformes), but probably larval lampridiforms.

Suborder Megalomycteroidei

Family Megalomycteridae

Four rare, little-known, deep-sea genera, probably larval lampridiforms.

ORDER ATHERINIFORMES

Premaxilla greatly expanded between maxilla and mandible, without crossed ligaments controlling the upper jaw, infraorbital bone series incomplete.

Suborder Exocoetoidei

Lateral line complete and low on the flank in marine forms, the lower pharyngeal bones are fused, no parietals, 9–15 branchiostegals. Found worldwide, but especially abundant in the Indo-Pacific.

Family Exocoetidae (halfbeaks and flying fishes)

Lower jaw often extended; snout not modified. Surface marine waters and freshwaters, worldwide; length to 45 cm (18 in.). Fossil half-beaks in the middle Eocene.

Family Belonidae (garfishes and needlefishes)

Snout bones sutured together, both jaws elongated into a strongly toothed beak. Mostly temperate and tropical marine; a few freshwater; length to 120 cm (almost 48 in.). Fossils in the Oligocene.

Family Scomberesocidae (sauries, skippers)

Snout and jaws as in Belonidae but feebly toothed; small finlets behind dorsal and anal fins. Inshore temperate and tropical marine waters; length to 35 cm (almost 14 in.). Fossils in the Miocene.

Suborder Cyprinodontoidei

Lateral line represented by pits on the flank, 4–7 branchiostegal bones. Families mostly distinguished by reproductive specializations.

Family Oryziatidae (medakas)

Most primitive cyprinodonts; a single genus in freshwaters and brackish waters in Indonesia.

Family Adrianichthyidae

Mouth and snout enlarged and shovellike. Two genera in lakes in Celebes; length 7–20 cm (2¾ to almost 8 in.). Fossils in Late Tertiary in Celebes.

Family Horaichthyidae

Small fishes with anal fin modified, in males, for clasping female in mating. One genus, freshwater, India.

Family Cyprinodontidae (killifishes or egg-laying topminnows)

Circumtropical and temperate marine and freshwater, many genera. Many popular aquarium fishes; length to 15 cm (6 in.). Fossils in the Oligocene.

Family Goodeidae (Mexican topminnows)

Live-bearing, but male lacks elaborate intromittent organ found in poeciliids. About 10 genera, in rivers draining the Mexican Plateau; length to about 10 cm (4 in.).

Family Jenynsiidae

Small fishes with asymmetrical genital organs; 1 genus; rivers of South America.

Family Anablepidae (four-eyed fishes)

Characterized by specialized eye structure (see above *Form and function*); 1 genus, 2 species; surface waters in rivers and estuaries of South America.

Family Poeciliidae (live bearers or viviparous topminnows)

Native to tropical and subtropical America but introduced elsewhere for mosquito control. Freshwaters and coastal marine waters. Length 1.5 to about 15 cm (over ½ to 6 in.). Family includes mollies (*Mollienesia*), guppies (*Lebistes*), swordtails (*Xiphophorus*), and many other popular aquarium fishes, as well as the mosquito fishes (*Gambusia*).

Suborder Atherinoidei

Lateral line variable; 5–7 branchiostegal bones; separate spinous dorsal fin.

Family Melanotaeniidae

Many species; freshwater bodies of New Guinea and Australia. Compressed, deep-bodied; pointed snout; 5–20 cm (2 to 8 in.).

Family Atherinidae (silversides)

Lateral line absent; pelvic fins midway along belly; length 7–70 cm (2¾ to 27½ in.). Coastal and freshwater, worldwide in warmer regions. Many genera. Fossils from middle Eocene.

Family Isonidae

Pectoral fins unusually high on body. Small marine fishes; Indian and Pacific Oceans. Two genera.

Families Phallostethidae and Neostethidae

Males with priapium, an organ derived from pectoral and pelvic girdles, functioning to clasp the female. Tiny fishes (3–5 cm [1 to 2 in.] long); confined to freshwaters and brackish waters in Thailand, Indonesia, and the Philippines.

Critical appraisal. The whale fishes (cetomimids, rondeletiids, barbourisiids) are often placed in a separate order Cetomimiformes, thought to be more primitive than Beryciformes, but their "primitive" features appear to be due only to degeneration. The stephanoberycids, melamphaeids, and gibberichthyids are usually placed in a suborder Stephanoberycoidei, all other beryciforms being placed in the Berycoidei, but the major phyletic cleft in Beryciformes seems to be between the holocentrids and the remainder, which form a related group. (C.P.)

Sticklebacks, tube snout, sea horses, and allies (Gasterosteiformes)

Gasterosteiformes is an order of fishes characterized generally by soft fin rays, pelvic fins located on the abdomen, an air bladder without a duct to the gut, and a primitive kidney. Gill structures are somewhat degenerate. Most species have bony rings around the body or ganoid (*i.e.*, thick, bony, enamelled, and diamond-shaped) plates rather than scales. Families within the order are Gasterosteidae (sticklebacks), Aulorhynchidae (tube snout), Indostomidae (indostomid), Aulostomidae (trumpet fishes), Fistulariidae (cornetfishes), Centriscidae (shrimpfishes), Macrorhamphosidae (snipefishes), Solenostomidae (ghost pipefishes), and Syngnathidae (pipefishes and sea horses).

Gasterosteiform fishes occur in both salt water and freshwater and are widely distributed. The smallest species are about three centimetres (about 1¼ inches) long, the largest about 200 centimetres (about 80 inches). They

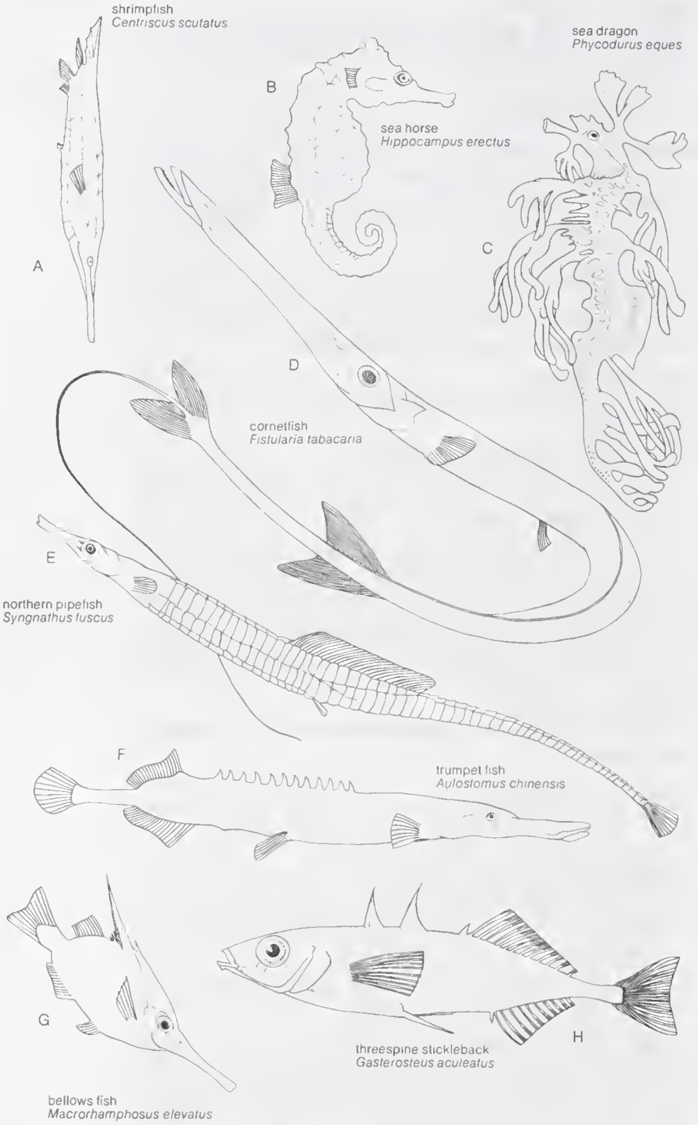


Figure 31: Representative gasterosteiform fishes.

Drawing by J Helmer based on (G, F) G Whitley and J. Allan, *The Sea-Horse and Its Relatives*, (D, E, H) A H Leim and W B Scott, *Fishes of the Atlantic Coast of Canada* (1966), Fisheries Research Board of Canada, reproduced by permission of Information Canada

are of limited economic importance, but many forms are popular aquarium fishes. Two families, Indostomidae and Aulorhynchidae, are represented by only one species each.

NATURAL HISTORY

Reproduction and life cycle. Except for sticklebacks, pipefishes, and sea horses, little is known of the life cycles of Gasterosteiformes. The male stickleback builds nests of plant materials cemented together with mucus secretions. The usually drab body hues of the male change, reflecting red, which is sexually attractive to the female. Male pipefishes and sea horses brood the eggs, deposited by the female within the male's brooding pouch. The brooding organ of the male sea dragon is a specialized area of soft skin beneath the tail. In some ghost pipefishes, eggs are fastened to the female on filaments of skin in pouches formed by specialized fins on the ventral, or lower, side. The dwarf sea horse, *Hippocampus zosterae,* breeds nine months of the year; eggs hatch after 10 days into miniatures of the adult. They mature in two or three months and live less than a year. Tube snouts deposit eggs in cavities of ascidia (primitive colonial chordates) or in masses of algae bound with threads secreted in a manner similar to that of sticklebacks. Egg clusters, often multiple, are cared for by the male. Snipefish eggs are enveloped in a mucilaginous substance from which the larvae are freed as development proceeds. Cornetfishes lay free pelagic (*i.e.,* drifting) eggs; thus, they do not receive parental care. The reproductive habits of shrimpfishes and trumpet fishes are unknown.

Brooding organ of the male

Ecology and behaviour. For defense, most gasterosteiforms assume a vertical position among grasses, gorgonians (*i.e.,* sea fans, a type of coral), and sea urchins. Such a posture serves to camouflage them; it also tends to present body spines or shields to predators normally oriented to the horizontal plane.

In most families, locomotion is by means of the caudal, or tail, fin. Snipefishes swim forward or backward with equal ease on the vertical plane and do not seek shelter among marine growths. The caudal fin is absent in sea horses. The coiled tail of the sea horse is used for gripping seaweed and other plants or objects. Propulsion is by means of the dorsal fin (*i.e.,* the large fin arising from the midline of the back). Tiny pectoral fins are used for steering. Vertical movement is by swimming. All fishes with an air bladder use it to some degree for vertical motion. With little effort the sea horse rises or settles to another depth by changing the air volume within the bladder.

With the exception of the snipefishes, most gasterosteiforms live among a wide variety of aquatic growths where they find food and safety and reproduce. Certain pipefishes and sticklebacks, in particular, are able to tolerate a wide range of salinity.

FORM AND FUNCTION

Sticklebacks are the most varied in form. The number of spines and bony plates is greatest in individuals living in the ocean. Each heavily armoured marine species is represented by half-mailed or naked (plateless) varieties in brackish or fresh waters. They are small scaleless fishes that grow to about 15 centimetres (six inches) in length. The short jaws are well armed with sharp teeth. The body is more nearly fusiform (*i.e.,* tapered at both ends) than are those of other members of the order. Body plates may be absent or may vary in number. The soft dorsal fin is preceded by from two to 11 free spines, each connected to the dorsal surface by its own triangular membrane. Pelvic fins are thoracic (*i.e.,* near the midsection) in position, each with a well-developed spine and one or two soft rays. The anal fin is preceded by a spine. The caudal fin is truncate (*i.e.,* abbreviated).

Sticklebacks

The body of the tube snout is elongated, slender, and cylindrical. It is tipped by a prolonged snout, the small, toothed mouth of which has a hinged upper jaw. The scaleless body is armoured with series of embedded bony plates. The first dorsal fin is represented by about 25 free spines; the rayed dorsal fin is far back on the body above the anal fin. Pectoral fins are broad and the caudal fin furcate (forked).

In the indostomids the elongated body is covered with bony rings as in pipefishes and sea horses. The small mouth is at the tip of the snout. The teeth are minute, the gills rather lobe shaped, and the eyes large. The anterior (*i.e.,* forward) dorsal fin consists of five isolated spines. Ventral fins (*i.e.,* paired fins arising from the sides of the belly) bear no spine, have four rays each, and are located not far behind the pectorals. The anal fin is below the soft dorsal fin, and the rounded caudal fin has a short peduncle, or stem.

Trumpet fishes, which seldom grow to more than 30 centimetres (one foot) in length, have an elongated, compressed, scaled body; the snout is prolonged into a rigid tubelike beak. The short, weak jaws have minute teeth. There are numerous dorsal spines. The ventral fins are abdominal; the caudal fin is truncate.

Cornetfishes, which grow to more than 180 centimetres (six feet) in length, are similar in structure to trumpet fishes; however, there are no scales. Instead, bony plates are embedded in the skin. Dorsal spines are absent, and the ventral fins are located in the abdominal region; each has a spine and four rays. Four anterior vertebrae are elongated. The backbone extends through the forked caudal fin as a long central filament.

Shrimpfishes, also known as razor fishes, are small, with toothless jaws at the end of a long snout. Scales are absent; the back is covered by transparent plates, forming the cuirass. Anteriorly the cuirass is affixed to the ribs;

posteriorly it extends beyond the displaced dorsal, caudal, and anal fins. The body is compressed to a sharp edge ventrally; hence the name razor fish.

In snipefishes the tubular snout has short jaws. The body form is variable, but all snipefishes tend to be short and deep and partly covered with the bony-plated cuirass, which is strengthened by its union with parts of the vertebrae. Areas lacking plates sometimes have scales. In addition to having several shorter spines, the dorsal fin has a very long, strong, serrated spine reaching nearly to the tip of the caudal fin. Each ventral fin has one spine and five rays. A long snout, and two posterior "handles" of spine and tail lends an appearance that is the basis for another common name, bellows fish.

Ghost pipefishes have a tubular snout tipped with a small mouth; the short body has spinous dorsal and ventral fins. Gills are reduced to lobe-shaped tufts attached to rudimentary gill arches. Bony plates unite to form rings. This supporting external framework has reduced the need for well-developed musculature. Ventral fins and a pointed caudal fin are large in proportion to the body size.

Pipefishes

Pipefishes are long and slender. The axis of the head is in line with that of the body, and the long snout is tipped with a small mouth. Bony rings replace scales. The dorsal and pectoral fins are spineless, and ventral fins are absent. Generally, the caudal fin is rounded and reduced, but it is effective in moving the fish rapidly through the water. The slender posterior body portion, though not truly prehensile (*i.e.*, capable of coiling and grasping), can be somewhat used in that manner. Sea horses are similar to pipefishes but differ in several important respects. The head is at an angle to the body proper. This, in addition to the shape of the head, creates a somewhat horselike appearance. The tail is prehensile and lacks a caudal fin.

EVOLUTION AND CLASSIFICATION

Paleontology. Gasterosteiformes appears to represent an early but highly specialized branch of the Acanthopterygii. Its evolution has been traced through limited paleontological data. Fossil sticklebacks occur in Miocene (7,000,000 to 26,000,000 years ago) strata and are moderately abundant in Tertiary (2,500,000 to 65,000,000 years ago) strata. Most of the order occurs in fossil remains in Eocene (38,000,000 to 54,000,000 years ago) and Oligocene (26,000,000 to 38,000,000 years ago) strata in the area of Monte Bolca near Verona, Italy. Deposits there (and in Sumatra) include an extinct family, Protosyngnathidae, related to the tube snouts. Trumpet-fish species, a closely related extinct scaleless family, Urosphenidae, and a small species of cornetfish have also been found there. Shrimpfishes, too, are within these strata, as well as in Oligocene deposits in various parts of Europe. Snipefishes are represented by an extinct genus, as are ghost pipefishes. Pipefishes occur in Miocene deposits in Sicily. Tertiary rocks contain *Syngnathus* and *Calamostoma*, pipefish forms that have a close relationship to true sea horses. Fossil sea horses are unknown.

Distinguishing taxonomic features. The Gasterosteiformes are classified mainly on the basis of general body form, the structure and distribution of scales or body plates, fin form and position, and the structure of the skeleton and its individual parts.

Annotated classification. The classification here is essentially that of P.H. Greenwood *et al.* (Bulletin of the American Museum of Natural History, no. 131, 1966).

ORDER GASTEROSTEIFORMES

Eocene to Recent. Frequently with strong spines in dorsal and pelvic fins, spines absent in some; snout often elongated; body often with dermal plates; 9 families, marine and freshwater, widely distributed. Length about 3–200 cm.

Suborder Gasterosteoidei

Family Gasterosteidae (sticklebacks)

Jaws short, armed with sharp teeth; body quite fusiform (tapered at both ends); body plates may be absent or may vary in number; body length to about 15 cm (6 in.); 11 species, fresh, brackish, and marine waters of Northern Hemisphere.

Family Aulorhynchidae (tube snout)

Body elongated, slender, and cylindrical; snout long, upper jaw hinged. One species, occurs in northeastern Pacific Ocean.

Family Indostomidae (indostomid)

Body elongated, covered with bony rings; teeth minute, gills lobe-shaped, eyes large. One species, *Indostomus paradoxus*, found in Lake Indawgyi in northern Burma.

Suborder Aulostomoidei

Family Aulostomidae (trumpet fishes)

Body elongated and compressed sideways; jaws short and weak, teeth minute; dorsal spines numerous; length to about 30 cm (12 in.); about 4 species, tropical seas.

Family Fistulariidae (cornetfishes)

Similar in appearance to Aulostomidae; no scales, bony plates imbedded in skin; dorsal spines absent; backbone extends through caudal fin as a central filament. Grow to more than 180 cm (71 in.); about 4 species, tropical seas.

Family Centriscidae (shrimpfishes)

Body small, jaws toothless, scales absent, back covered by transparent plates; 4 species, shallow waters of Indian and Pacific Oceans.

Suborder Syngnathoidei

Family Macrorhamphosidae (snipefishes)

Snout tubular, jaws short, body rather short and deep; in profile shaped like bellows; 11 species, temperate and tropical seas.

Family Solenostomidae (ghost pipefishes)

Snout tubular, mouth small; body short, with spiny dorsal and ventral fins; bony plates united to form body rings; about 5 species, tropical Indo-Pacific waters.

Family Syngnathidae (pipefishes and sea horses)

Pipefishes long and slender, snout tipped with small mouth; dorsal and pectoral fins spineless, ventral fins absent. Sea horses with head bent downward in horselike relation to body; tail prehensile; bony rings instead of scales; about 24 species of sea horses, widely distributed, marine; about 150 species of pipefishes widely distributed in shallow tropical seas.

Critical appraisal. The American Fisheries Society (Special Publication No. 6, 1970) arranges the Gasterosteiformes as in Greenwood *et. al.*, above, with the following exceptions: the taxonomic level of suborder is not used; family Gasterosteidae is expanded to include family Aulorhynchidae; family Centriscidae is expanded to include family Macrorhamphosidae.

Of the several families considered in this article, the Indostomidae are the least known. The species of this family appear to be intermediate between Gasterosteidae and Aulostomidae.

A tenth family, Pegasidae, is tentatively placed between Gasterosteidae and Syngnathidae. Some authorities list them separately as Pegasiformes. These small fishes are the so-called sea moths, found in Asiatic seas. The toothless mouth is not terminal but lies under the head and is overhung by a snout, or rostrum, often adorned with spines. The body is protected by knobby armoured plates, with the posterior portion rather elongated, square to rectangular in cross section and bearing a small dorsal fin. Spines along either side of this region may be absent or developed to varying degrees. Pectoral fins form expansive fans on either side. Ventral fins are reduced to a few fingerlike rays used for crawling on the bottom. The pegasids have no air bladder. Swimming ability for more than short distances is poor. Fossils of Pegasidae are unknown. (W.Z.)

Scorpion fishes, rockfishes, gurnards, and allies (Scorpaeniformes)

The scorpaeniform, or mail-cheeked fishes, are widespread throughout the oceans of the world. The group is believed to have originated in warm marine waters but has invaded temperate and even Arctic and Antarctic seas as well as freshwaters of the Northern Hemisphere. The mail-cheeked fishes are a highly successful biological group, occurring in the sea from the midlittoral (coastal) zone down to depths of at least 4,000 metres (about 13,000 feet). They inhabit some deep freshwater lakes but are more abundant in cold streams and rivers. The order is often divided into six suborders, only two of which have more than one family; these are the Scorpaenoidei (three families) and the Cottoidei (seven families). The best known groups are the scorpion fishes and rockfishes (family Scorpaenidae), gurnards or sea robins (Triglidae), flatheads

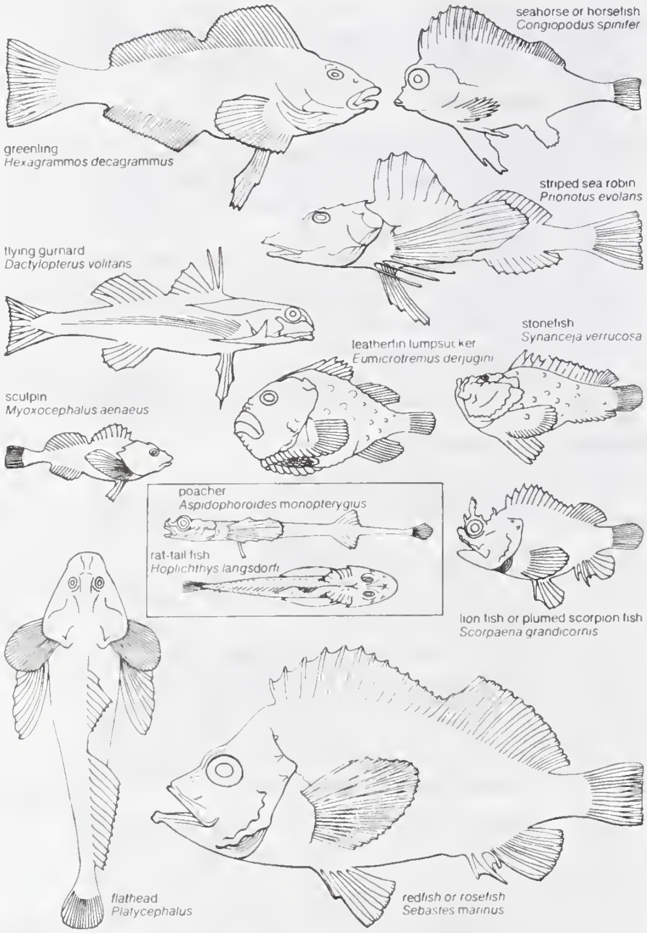


Figure 32: Body plans of representative Scorpaeniformes.

From (*Hexagrammos, Platycephalus*) Yaichiro Okada, *Fishes of Japan;* (*Congiopodus*) J L B Smith, *Sea Fishes of Southern Africa*, (*Prionotus, Dactylopterus, Aspidophoroides, Scorpaene, Sebastes*) D S Jordan, *A Guide to the Study of Fishes*, copyrighf 1905 by Holt, Rinehart and Winston, Inc., reprinted by permission of Holt, Rinehart and Winston, Inc , (*Eumicrotremus*) adapted from original drawing by D R Harriott in A H Leim and W B Scott, *Fishes of the Atlantic Coast of Caneda*, Fisheries Research Board of Canada Bulletin 155, (*Myoxocephalus*) reprinted by permission of G P Putnam's Sons from *Field Book of Marine Fishes of the Atlentic Coast* by Charles M Breder, Jr , copyrighf © 1929 by Charles M Breder, Jr , renewed 1957 by Charles M Breder, Jr

(Platycephalidae), and sculpins (Cottidae). The flying gurnards (Dactylopteridae), considered by some authorities to belong in this order but more often separated in the order Dactylopteriformes, are treated here for convenience.

GENERAL FEATURES

Many members are locally important commercial fish. Thus the redfishes of the genera *Sebastes* and *Sebastodes* of the North Atlantic and Pacific have considerable value to the fishing industries of Europe, Russia, and North America; the flatheads are exploited in a wide area of the Indo-Pacific region, and greenlings (Hexagrammidae) are of commercial importance in the northwestern Pacific. In general, the fishery value of the group as a whole has a greater potential than is shown by the present actual utilization by man.

Members of the order Scorpaeniformes are not large fishes. Some of the deepwater species, such as the redfishes, grow to a length of 90 centimetres (three feet), but the majority attain a maximum length of around 30 centimetres (one foot). Externally, scorpaeniforms vary greatly; most are like the Perciformes in general appearance—*i.e.,* they are typical, scaled, spiny-rayed fishes—but the lumpfishes (Cyclopteridae) among them are obese, often jellylike, and usually scaleless and lack sharp fin spines. Body armour is often well developed, however, and most scorpaeniforms are well equipped with spines.

NATURAL HISTORY

Ecology. The greatest diversity of scorpaeniform fishes, especially among members of the family Scorpaenidae, is found in warm tropical seas. The order is particularly well represented in the shallow waters of the continental shelf. In the tropics many scorpaenids are found in association with coral; elsewhere most are found on rocky or rough bottoms. In both cases they are highly camouflaged to match their backgrounds. The sculpins, flatheads, and scorpaenids, although not fast or powerful swimmers, feed on many more active smaller fishes and crustaceans, usually capturing prey by a swift pounce using the large pectoral fins as auxiliary power units.

The dangerously venomous stonefishes

Cryptic (concealing) coloration is probably best exemplified by the stonefishes (*Synanceja*), inhabitants of shallow waters, including estuaries, mud flats, coral reefs, and coral sand pools of much of the tropical Indo-Pacific. Looking like a lump of eroded coral or rock, the stonefish's body is concealed by the cragginess of its outline and by its coloration, which exactly matches that of the background. A stonefish is perfectly hidden and makes no movement until prodded with a stick or, more often, until it is stepped on, at which time it erects its dorsal fin spines. Each spine has a pair of large gray-brown fusiform venom glands, one on each side. The spines inflict puncture wounds into which a considerable quantity of venom is injected through a channel on each side of the spine. Intense pain at the site of the puncture is instantaneous and radiates within minutes to involve the whole of the affected limb. Death occasionally occurs, and secondary infections are common.

Many other scorpaeniforms can inflict severe wounds with their fin spines or head spines, but relatively few species are equipped with venom glands. One group of venomous species includes the turkey fishes (*Pterois* and related genera), also known as lion fishes or fire-fishes. Widespread in tropical Indo-Pacific waters, they are beautifully and boldly coloured, with patterns of contrasting stripes on the head and body that are specific for individual species and extremely long dorsal and pectoral fins. All of these fishes have long, needlelike dorsal spines with glandular venom-producing tissue and shallow channels capable of inflicting very painful but rarely fatal punctures. The bold and distinctive colouring of the turkey fish is clearly a warning, for, unlike most scorpaenids, it does not hide but boldly swims in open water around the coral heads. If disturbed, the turkey fish displays by spreading its fins to their fullest extent, rotating until it assumes a position, often head down, with its dorsal spines pointing toward the intruder. If an attacker is not intimidated by this display, the turkey fish moves toward the attacker with its dorsal spines erect.

The flatheads (Platycephalidae) are found in the same oceans as the scorpaenids but mainly in sandy, muddy, or estuarine areas. Their greatly flattened bodies are clearly an adaptation to bottom life; indeed, they bury themselves on the bottom, leaving only the eyes exposed. Many species feed mainly on small fishes, but others, like the dusky flathead (*Platycephalus fuscus*), the largest and commercially most valuable of the Australian flatheads,

From B W Halsted, *Poisonous and Venomous Marine Animals of the World;* U.S Detense Department

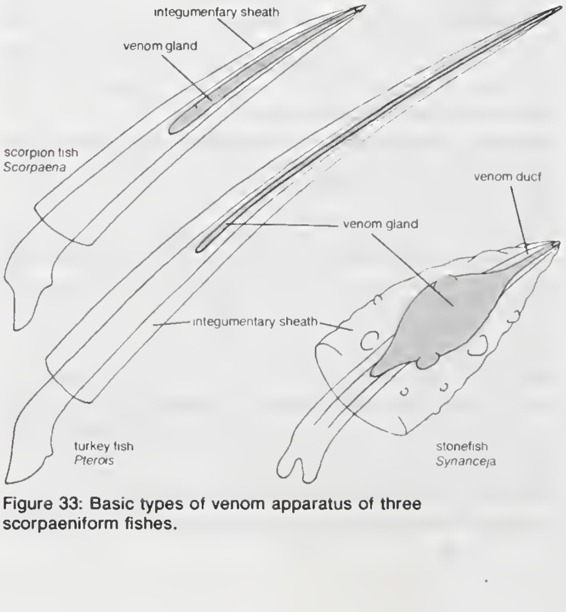


Figure 33: Basic types of venom apparatus of three scorpaeniform fishes.

have a varied diet of fishes, mollusks, crustaceans, and marine worms.

Adaptations for bottom living in sea robins

The sea robins (Triglidae) are bottom-living fishes of wide distribution. The lower two or three pectoral fin rays, which are long, thickened, and detached from the remainder of the fin, form organs of taste and touch and are used for locomotion. These rays are very mobile, and an active sea robin can move slowly along the bottom apparently supported on the rays, which continuously explore the ground ahead and on either side. The diet of most of these fish consists of crustaceans, mollusks, and other fishes, many of which burrow in the seabed.

The most abundant littoral (shore) scorpaeniforms are the sculpins (Cottidae), significantly the only group found in fresh water, other than the closely related families Cottocomephoridae and Comephoridae. Some members of the families Scorpaenidae and Cyclopteridae are also littoral fishes. The littoral sculpins are generally small, inhabiting densely weeded pools or crevices in rocks. Both the sculpins and cyclopterids found along the shore are strongly thigmotactic (attracted to surfaces), pressing as much of their bodies to the surface as possible. The cyclopterids have well-developed sucker disks, which are derived from the pelvic fin complex. The suckers, which are effective in resisting wave action, are capable of exerting considerable force; in one instance, a force of 13.3 kilograms (approximately 29 pounds 5 ounces) was required to break the hold of an adult lumpsucker (*Cyclopterus lumpus*). The European littoral sea snail (*Liparis montagui*) can vary the suction exerted by its sucker as necessary to adjust for the speed of passing water currents.

The cyclopterids have adopted a wide variety of lifestyles in addition to the littoral habit. The genus *Nectoliparis* is pelagic (*i.e.*, inhabiting open water); members of the genera *Paraliparis* and *Rhodichthys*, of the North Pacific and Arctic Oceans, are bathypelagic, at least for a large part of their lives. In fishes found at depths of 2,400 metres (7,900 feet), the pelvic sucker disk is completely absent. In the semitransparent but beautifully pink-tinged species of the genus *Careproctus*, found in deep, cold polar waters, the pelvic sucker is greatly reduced in size and presumably in efficiency.

In contrast to the cyclopterids, the greenlings are pelagic fishes that adopt a benthic (bottom) life only during the spawning season. One of the best-known members, the Atka mackerel (*Pleurogrammus monopterygius*), which is common in the North Pacific and has considerable sporting and commercial fishing value, spends the major part of its life in the open sea. The related yellow-fish (*P. azonus*) has been observed in the upper layers of the ocean in calm weather and is usually captured in purse seines. At night it descends to the bottom.

The scorpaeniforms have adapted particularly well to fresh water. The members of the sculpin family (Cottidae) are widely distributed in the Northern Hemisphere, reaching their greatest diversity in North America and decreasing in number westward through the Eurasian landmass. In extreme western Europe there is only one species, the miller's thumb (*Cottus gobio*). Two endemic forms, *C. kneri* and *C. kessleri*, both of some commercial importance, are found in Lake Baikal, Russia, and its tributary rivers. These species share Lake Baikal with a number of other species belonging to the related families Cottocomephoridae and Comephoridae. The sculpins in Lake Baikal have become adapted to exploit all of the living space offered by this inland sea. The family Cottocomephoridae is divided into eight genera and numerous species. Although many of the benthic species are restricted to a particular type of bottom and are found only within a certain depth range, most migrate into coastal waters in the spring and remain there during the warm season. Some species, however, remain in deep water all year; others, which are primarily pelagic fishes, use the bottom only to spawn.

Diversity of sculpins in Lake Baikal

The two members of the family Comephoridae, called Baikal cods (*Comephorus baicalensis* and *C. dybowskii*), are pelagic fishes, the latter living at depths to 1,000 metres (more than 3,000 feet). The feeding habits of these Baikal cottoid fishes all exploit potential food resources; the pelagic species feed mainly on various pelagic crustaceans and make daily vertical migrations accompanying their prey, and the benthic forms feed chiefly on certain species of benthic copepod crustaceans. The diet of one inshore species, *Batrachocottus nikolskii*, however, is mainly chironomid and caddisfly larvae.

Reproduction. The mail-cheeked fishes are highly variable in their mode of reproduction. Some of the methods used by them to reproduce are noted below. In the Comephoridae there is a remarkable imbalance between the numbers of each sex, the proportion of males in the total population being as low as 3 or 4 percent. The biological basis for this imbalance is unknown. Members of this family are viviparous (live-bearing). The females come near the surface to give birth to their young; the males remain at their normal depths. By contrast, the remaining cottoids are oviparous (egg-laying), including the Cottocomephoridae of Lake Baikal. The females of most of the latter family deposit their eggs in shallow coastal water, then leave the males to guard them until they hatch. This is also the general rule among the sculpins, in which the males guard the eggs. In some species the eggs are shed loosely and adhere to the bottom, but little reliable evidence is available concerning the breeding habits of most cottoid fishes. The northern Atlantic short-horned sculpin, or bullrout (*Myoxocephalus scorpius*), is known to build a rudimentary nest guarded by the male, as does the freshwater European miller's thumb. The males of these and other cottids have a well-developed structure called a urogenital papilla, which some authorities have suggested is used to introduce sperm into the female. Many cottoid species develop pronounced breeding coloration with sexual differences that apparently aid in recognition between the sexes and in territorial behaviour.

Specialized behaviour

Some members of the family Cyclopteridae build nests that are guarded by the male. The familiar lumpsucker, or sea hen (*Cyclopterus lumpus*), common on both sides of the North Atlantic, spawns along the coast in the winter. At least some of the inshore species, such as the striped sea snail (*Liparis liparis*), which has a distribution similar to that of the lumpfish, deposit their spawn in clumps on hydroids (*e.g.*, sea moss) and seaweeds, but there is no evidence of parental care. The females of some North Pacific cyclopterids (*e.g.*, *Careproctus sinensis*) have a long specialized structure (ovipositor) by which they lay their eggs under the shell of the Kamchatka crab. In general, however, little is known about the breeding biology of these fish.

In contrast to the rather specialized reproductive behaviour of the sculpins and the sea snails, the sea robins produce eggs that are simply shed in batches in the open sea. So far as is known, no special breeding behaviour accompanies spawning except that these noisy fishes become increasingly loquacious during the spawning season. The members of the family Hexagrammidae (the Atka mackerel, for example) deposit one clump of eggs, often on algae in shallow water on stony bottoms; some species, however, like the lingcod (*Ophiodon elongatus*), care for their egg masses during incubation. The sea poachers, or pogges (Agonidae), lay relatively few eggs, often hiding them away in crevices. The eggs are relatively large, 1.5–1.9 millimetres (roughly $^1/_{16}$ inch) in diameter in *Agonus decagonus*, a species found in the extreme North Atlantic. The European hook-nose (*Agonus cataphractus*) lays up to 2,400 eggs inside the hollow rhizoid ("stalk") of the kelp *Laminaria* in a compact, membrane-covered mass. Incubation is prolonged, possibly as long as 12 months.

Scorpion fishes of the family Aploactinidae similarly shed their eggs in the open sea. Members of the scorpaenid subfamily Scorpaeninae extrude eggs in gelatinous balloon-shaped masses; those in the subfamily Sebastinae have internal fertilization and are viviparous. The three groups represent the principal evolutionary stem groups of the scorpion fishes. In the North Atlantic redfish (*Sebastes marinus*) fertilization is internal, and the eggs develop within the oviduct of the mother. Fertilization usually takes place during February, after which the females form shoals and migrate to spots where warm bottom currents pass. The female can be said to be a living incubator; in such fishes, which live in cold northern seas, it is clearly advantageous to carry the developing young to an area in which more favourable conditions prevail. The young at birth are very immature, nevertheless, and brood size in the redfish is relatively large (up to 360,000); the larvae must survive a lengthy planktonic life. A smaller, shallower water redfish, the Norway haddock (*Sebastes viviparus*) produces much smaller broods, with brood sizes ranging from 12,000 to 30,000 young. The scorpaeniforms are distinguishable among viviparous teleosts (advanced bony fishes) by their comparatively high fecundity; comparison with many other marine fishes, single individuals of which produce millions of freely shed eggs, however, illustrates the relative advantage, at least in numbers, of bearing living young over laying eggs.

Differences in numbers of offspring

The North Pacific redfishes or rockfishes (*e.g., Sebastodes, Sebastiscus,* and *Hosukius*) closely resemble the North Atlantic sebastine species in their reproductive biology; all species studied have been found to have relatively large brood sizes. The sebastine rosefishes (*Helicolenus*), found in both the northern Atlantic and Pacific Oceans, have morphological affinities with the subfamily Scorpaeninae. Studies of their reproductive biology have shown that the sebastine rosefishes have intraovarian embryos embedded in a gelatinous matrix, and they thus appear to combine sebastine viviparity with scorpaenine egg masses.

Sound production. Since the time of Aristotle, the sea robins have been known as sound-producing fishes, and their sonic performances and mechanisms are well known. They have a large swim bladder loosely attached to the dorsal wall of the body cavity; the swim bladder is vibrated by lateral muscles in which the striated fibres run at right angles to the muscles' length. The sea robin of the North Amerian Atlantic coast (*Prionotus carolinus*) produces single vibrant barks and growls, as wells as series of rapid clucks with very little provocation. Some of the sculpins (Cottidae) produce dull groans and growls; it is believed that these sounds are mechanical in origin, arising from contractions of the muscles that produce periodic movements of the pectoral girdle. Flying gurnards (Dactylopteridae) are similar to the triglids in their sonic mechanism and sound production capacity.

FORM AND FUNCTION

General features. Many mail-cheeked fishes, such as the rockfishes, have simple fusiform (spindle-shaped) body plans, but others, such as some sea poachers, are extremely slender. Most scorpaenids, triglids, and cottids have two dorsal fins (sometimes joined), the forward one supported by stiff spines. In general, the dorsal, anal, and pectoral fins are large, sometimes strikingly so, but in some groups (*e.g.,* cyclopterids) the spinous dorsal fin is reduced or absent.

The uniting feature of the order

The one feature diagnostic of the order is the presence of a bony bar, or stay, beneath the eye, an extension of the second infraorbital bone. It is small and inconspicuous in some primitive scorpaeniforms and secondarily reduced in some specialized forms; in others, however, it is readily evident. Often it has become externally prominent, bearing protrusions or spines, or has expanded and fused other cranial bones bones to form a hard armour.

Camouflage and coloration. Many scorpaeniform fishes, such as scorpion fishes, rockfishes, and sculpins, which live on coral or rocky bottoms, possess a remarkable degree of cryptic (concealing) coloration and shape. Numerous fleshy lappets adorn the head, fin membranes, and body scales, rendering the fish virtually invisible against a background of rocks covered with marine organisms. The effectiveness of this camouflage cannot be appreciated unless the fish is viewed in its natural habitat.

Some members of the order, such as the sea robins, are notable for brilliant colours, especially reds. The large pectoral fins are often strikingly coloured; in the European tub gurnard (*Trigla lucerna*) they have spots of bright green and peacock blue. The spots on the pectorals of the flying gurnards resemble eyes and apparently function to startle and frighten potential predators when the fins are suddenly spread. The brightly coloured fins of some sea robins may function in a similar manner.

CLASSIFICATION

Annotated classification. The following classification is modified from that suggested by British ichthyologist P.H. Greenwood and the Americans S.H. Weitzman, D.E. Rosen, and G.S. Myers.

ORDER SCORPAENIFORMES (mail-cheeked fishes)
Spiny-finned fishes generally with stout bodies. Second infraorbital bone united with the preopercular to form a rigid stay across the cheek. Pelvic fins thoracic in location (sometimes modified into sucker disks), the bones directly attached to cleithra (bones like the collarbones of higher vertebrates).

Suborder Scorpaenoidei
Moderate-sized fishes with 24 to 40 vertebrae. Anterior ribs absent or sessile (rigidly attached). Numerous species.

Family Scorpaenidae (scorpion fishes and redfishes)
Paleocene to present. Marine fishes widely distributed in tropical, temperate, and northern seas. Perchlike appearance, dorsal fin spines long and numerous; head spiny, body scaly. Locally important food fishes, some with venom glands on fin spines. Size to around 100 cm (39 in.).

Family Triglidae (gurnards and sea robins)
Upper Eocene to present. Marine fishes of warm and temperate seas. Characterized by rather slender form, body with small scales or bony plates; head heavily armoured. Lower pectoral fin rays separate, forming a tactile organ. Locally exploited by man as food. Size to around 70 cm (28 in.).

Family Synancejidae (stonefish)
Tropical Indo-Pacific. Scaleless, body covered with warty tubercles. Venom glands on fin spines; dangerously venomous. Size to around 60 cm (24 in.).

Families Aploactinidae, Pataecidae, and Caracanthidae
Tropical Pacific, often in coral. Small scaleless fishes (except Caracanthidae, which have dense dermal papillae). Size to about 25 cm (10 in.).

Suborder Hexagrammoidei
Moderate-sized, slender-bodied fishes. Vertebrae 42–64. Ribs attached to strong parapophyses (projections of vertebrae). Few species, midwater and benthic.

Families Hexagrammidae (greenlings), *Anoplopomatidae* (skilfish), *and Zaniolepididae* (comb fishes)
Northern Pacific. Small scales, long dorsal fins, spines on the head few, powerful teeth in jaws. Locally important food fishes, some with sporting value. Size of most to about 45 cm (18 in.), some longer.

Suborder Platycephaloidei
Moderate-sized with head and anterior part of body strongly flattened. Vertebrae about 27. No air bladder.

Family Platycephalidae (flatheads)
Head and body flattened anteriorly. Marine; usually buried in soft bottom, some forms on coral; Indo-Pacific and tropical eastern Atlantic. Important commercial fishes in Southeast Asia, tropical Australia, and elsewhere. Size to 130 cm (52 in.) and 15 kilograms (33 pounds).

Suborder Hoplichthyoidei
Small fishes, with very depressed bodies. Scaleless, but body with bony plates. Head with heavy spiny ridges.

Family Hoplichthyidae
Found in moderately deep water in Indo-Pacific region. Size to 25 cm (10 in.).

Suborder Congiopodoidei
Moderate-sized fishes with angular bodies and well-developed dorsal fin spines. Scaleless, but sometimes rough skins.

Family Congiopodidae (horse fishes)
In moderately deep cold waters of Southern Hemisphere, off South America, Australia, and South Africa. Size to 75 cm (30 in.).

Suborder Cottoidei
Small to moderate fishes. Marine, from temperate to polar seas, and freshwater in Northern Hemisphere. Mostly without scales; many with spiny skins, others with bony plates. Many species.

Family Cottidae (sculpins and bullheads)
Oligocene to present. Generally large headed, with well-developed head spines. Marine and freshwater of the Northern Hemisphere (1 genus said to occur in the Tasman Sea). Mostly small, some up to 60 cm (24 in.).

Families Cottocomephoridae and Comephoridae
Similar to cottids but postcleithral bones absent or rudimentary. Freshwater, endemic to Lake Baikal, Russia. Size to about 16 cm (6 in.).

Family Icelidae (two-horn sculpins)

Head large, laterally compressed, and with small spines. Lateral line and dorsal fin base with scutes (plates). Small benthic fishes mainly of North Pacific, a few in the North Atlantic. Size to 25 cm (10 in.).

Family Normanichthyidae

Head and body with ctenoid (fringed) scales; head withoui spines. Two well-separated dorsal fins, soft fin rays branched. Marine waters off Chile. Size to 11 cm (4 in.). Probably does not belong in order Scorpaeniformes.

Family Cottunculidae

Body covered in loose skin covered wiih bony tubercles. Head large, lacking spines. Deep waters of the Arctic and North Atlantic oceans, Indian Ocean off South Africa, and Flores Sea, Indonesia. Size to 20 cm (8 in.).

Family Agonidae (poachers and pogges)

Eocene to present. Body covered in hard armour of large scutes. One or two dorsal fins. Teeth minute. Small, benthic, coastal fishes of northern Pacific, Atlantic, and Arctic oceans and Antarctic waters. Size to 30 cm (12 in.).

Family Cyclopteridae (lumpfishes and sea snails)

Body short, ihick, tadpole-shaped. Skin thick, naked or with bony tubercles or small thorns. Two dorsal fins, the first often minute or modified. Pelvic fins forming a sucker disk or absent. Marine, from littoral to abyssal depths (4,000 metres) in northern Atlantic and Pacific oceans, Arctic and Antarctic waters. Size to 60 cm (24 in.); 5.5 kg (12 lb).

ORDER DACTYLOPTERIFORMES

Resemble Triglidae. Head covered by bony plates ihat are expanded into huge shields. First infraorbital bone connected to preoperculum. Pectoral rays long, numerous, and brightly coloured.

Family Dactylopteridae (flying gurnards)

Tropical and warm temperate regions of Atlantic and Indo-Pacific oceans. Few species. Size to 50 cm (20 in.).

Critical appraisal. The classification of the scorpaeniforms cannot be said to have approached a final synthesis. It has been suggested that they are an aggregation of three distinct evolutionary lines, the two dominant elements being the scorpaenid and cottid-hexagrammid lines, the third, and minor, group being the anoplopomatids. On the other hand, British ichthyologist P.H. Greenwood and colleagues have pointed out that all members of the order, as recognized here, share a distinctive type of caudal (tail) skeleton. The order Scorpaeniformes is clearly related to Perciformes within a superorder Acanthopterygii.

The systematic positions of some groups remain open to doubt. The flying gurnards have been placed by some workers in the Scorpaeniformes, and, indeed, the morphological and biological resemblance of the flying gurnards to members of the Triglidae suggests that such placement best expresses their phyletic affinities. (A.Wh.)

Perches, tunas, marlins, and allies (Perciformes)

The order Perciformes, the perchlike fishes, is the largest group of fishes in the world today, comprising more than 6,000 species that have been classified into about 150 families. Perciform fishes occur in abundance in both marine and freshwater areas of the world, ranging from shallow freshwater ponds to depths of more than 2,300 metres (7,500 feet) in the oceans. Most perciforms are marine fishes, generally found along coastal areas of tropical and temperate regions of the world. The order includes many of the world's most important food and game fishes, such as tunas, mackerels, bonitos, and skipjacks (family Scombridae), billfishes and marlins (Istiophoridae), swordfish (Xiphiidae), sea basses (Serranidae), and carangids (Carangidae), a large family that includes pompanos, jacks, cavallas, and scads. The freshwater food and sport fishes of the perciform order include the sunfishes (Centrarchidae) and the perches and walleyes (Percidae). Many perciforms are popular aquarium fishes.

GENERAL FEATURES

Size range. Perciform fishes vary greatly in size, ranging from the tiny freshwater goby *Pandaka pygmaea* (Gobiidae) of the Philippines, which is fully grown at about 1.2 centimetres (less than one-half inch) in length, to the black marlin (*Makaira indica*), swordfish (*Xiphias gladius*), and bluefin tuna (*Thunnus thynnus*), which attain lengths of about 3.3 metres (11 feet). The bluefin tuna and the Indo-Pacific black marlin have been known to exceed 680 kilograms (1,500 pounds) in body weight. Generally, most percoid fishes fall within the range of 30 to 250 centimetres in length.

Distribution. Perciform fishes occur worldwide and are clearly a highly successful group. The coral reefs of tropical seas abound with colourful perciforms, including such species as wrasses, butterfly fishes, gobies, damselfishes, blennies, and cardinal fishes. The perciform order comprises a large part of the fauna of the Indo–West Pacific region, which is probably the world's richest in the variety of its fish fauna. Of the Antarctic fish fauna, approximately 75 percent belong to the order Perciformes. These cold-water perciforms include the icefishes (family Channichthyidae [Chaenichthyidae]), known for their "bloodless" appearance, which results from the lack or near lack of red blood cells and blood pigments. Freshwater perciforms include the cichlids (family Cichlidae), which occur naturally in India, Africa, South America, and parts of southern North America; these fishes also have been introduced elsewhere. The perch and sunfish families are found in North America and Europe, and the European perch (*Perca fluviatilis*) occurs well north in Siberia.

Drawing by G Christman

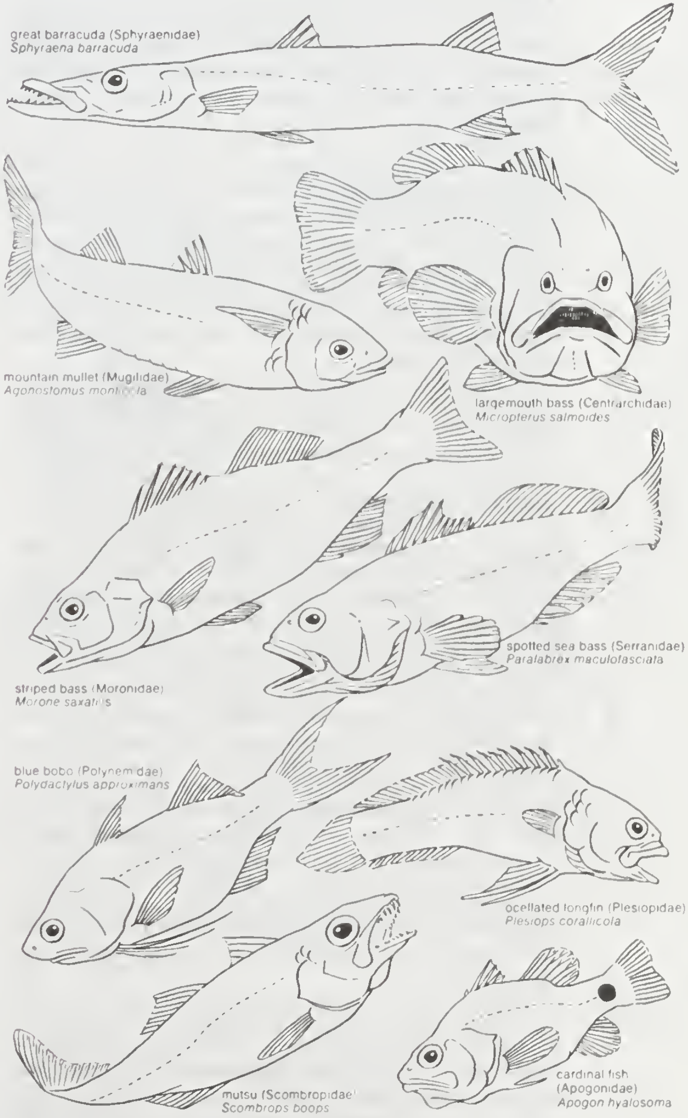


Figure 34: Representative perciforms of the families Sphyraenidae, Mugilidae, Polynemidae, Scombropidae, Apogonidae, Moronidae, Serranidae, Plesiopidae, and Centrarchidae.

IMPORTANCE

Use as food. Since early times, the rivers and oceans have provided man with food; fishing was one of man's earliest means for securing food. Archaeological findings among shell mounds of Scotland indicate that the sea bream (family Sparidae) formed part of the diet of early man. The Nile perches (family Latidae) have been found as mummies in ancient tombs in Egypt. The goatfishes (family Mullidae) appear in ancient Roman archives as one of the most highly valued food fishes, and in Japan the goatfish holds a good market and is eaten raw as *sashimi* or in the form of dried fish cakes known as *kamaboko.* In Japanese art through the ages, the fish god is shown with the "king of sea fish" under one arm; this highly valued food fish is the porgy *Chrysophrys major* (family Sparidae). A Japanese New Year's dinner usually includes *buriko,* the eggs of the sandfish *Arctoscopus japonicus* (family Trichodontidae).

The perciform fishes play an important part in commercial fisheries all over the world. Isinglass, which is used in the production of jellies and also in the process of clarification of wine and beer, is obtained from fishes that include the drums (family Sciaenidae) and the threadfins (family Polynemidae). The skin of the wolffishes (family Anarhichadidae) provides a leather of fair quality. The guanin present in the skin of the Japanese cutlass fish (*Trichiurus;* Trichiuridae) is used in the manufacture of artificial pearls in Japan.

Breeding and cultivation of perciforms has been successful in many parts of the world. The African mouthbreeder (*Tilapia macrocephala,* Cichlidae) has been successfully introduced in many areas and is valued for its rapid rate of reproduction and growth, providing a source of low-cost protein.

Aquarium fishes. Colourful and interesting perciforms are kept for aesthetic reasons by aquarists, augmenting an industry partially supported by fishes of other orders. Popular aquarium fishes of the perciform order include cichlids, butterfly fishes (Chaetodontidae), angelfishes (Pomacanthidae), labyrinth fishes (suborder Anabantoidei) such as the Siamese fighting fish (*Betta splendens*) and the kissing gourami (*Helostoma temmincki*), and various gobies (Gobiidae), blennies, and blennylike fishes of the suborder Blennioidei.

The freshwater angelfish *Pterophyllum scalare* and the discus (*Symphysodon discus*) are among the most popular aquarium fishes for breeding because of their remarkable means of feeding their young on the mucous secretions of their bodies.

Danger to human life. A few of the perciforms are known to be harmful to man. Swimmers have been attacked by the barracuda (*Sphyraena*), which is a voracious fish reaching nearly two metres (six feet) in length. Perciforms possessing venom glands are also considered dangerous fishes. The dorsal spine of the weever fishes (Trachinidae) has a grooved structure containing a venom gland; in addition, there is also a stinger located on the opercular (gill cover) structure. Both the stinger and the dorsal spine can be extremely painful if stepped on in shallow waters. Similar venom-bearing structures are found in the dragonets (Callionymidae) and surgeonfishes. The venomous spines in the surgeonfish are located on either side of the caudal peduncle (the narrow stalk just in front of the tail). Especially well armed are the electric stargazers (*Astroscopus,* Uranoscopidae), which are capable of discharging up to 50 volts of electricity from the modified muscle tissue just posterior to the eyes; in addition, they possess a venom spine just above the pectoral fins. The venom from uranoscopids has been known to cause death in man.

Poisonous perciforms

Ciguatera fish poisoning has been attributed to some perciforms that are otherwise considered to be excellent food fishes. Among these are certain carangids, snappers, barracudas (Sphyraenidae), surgeonfishes (Acanthuridae), groupers, and porgies. A species completely edible in one area may be poisonous in an area just a few hundred miles away. This curious phenomenon has not yet been fully explained, although it has been suggested that the source of poisoning may be a toxic form of blue-green alga passed up the food chain and thus present in the food of toxic species.

NATURAL HISTORY

Life history. Many perciforms live out their whole lives in small areas, but others, especially open-ocean (pelagic) species, perform extensive migrations, about which much remains to be learned. Some marine serranids, however, are anadromous (*i.e.,* entering fresh or brackish water to spawn); some freshwater perciforms, such as certain species of gobies, enter the sea to spawn (catadromous). Tuna (Scombridae) may travel across the entire Pacific Ocean from the California coast to Japan or the reverse. Spawning in perciforms generally takes place in shallow coastal areas or in rivers and ponds among rocks, seaweeds, and aquatic plants. *Paraclinus marmoratus,* a clinid blenny, is known to lay eggs at times in the lumen (cavity) of a living sponge.

Breeding behaviour among fishes of the order Perciformes is diverse. Pairing of male and female is common, although a single female may pair with more than one male, as among certain serranids, perches, and cichlids. The sexes are usually distinct, but hermaphroditism (presence of functional male and female organs in a single individual) normally occurs among certain sea basses and porgies. The young of the black sea bass (*Centropristis striata*) are mostly females with normal egg-laying functions; after five years, however, some of these females transform into functional males. About 11 species of sparids have been found to display hermaphroditism at one time or throughout their lifetime.

Differences in appearance between sexes

Characteristic differences usually exist, especially during the breeding season, between sexes regarding colour, size, markings, or structure. The male is generally smaller in size (some exceptions are found in sunfishes, gobies, and

Drawing by G. Christman

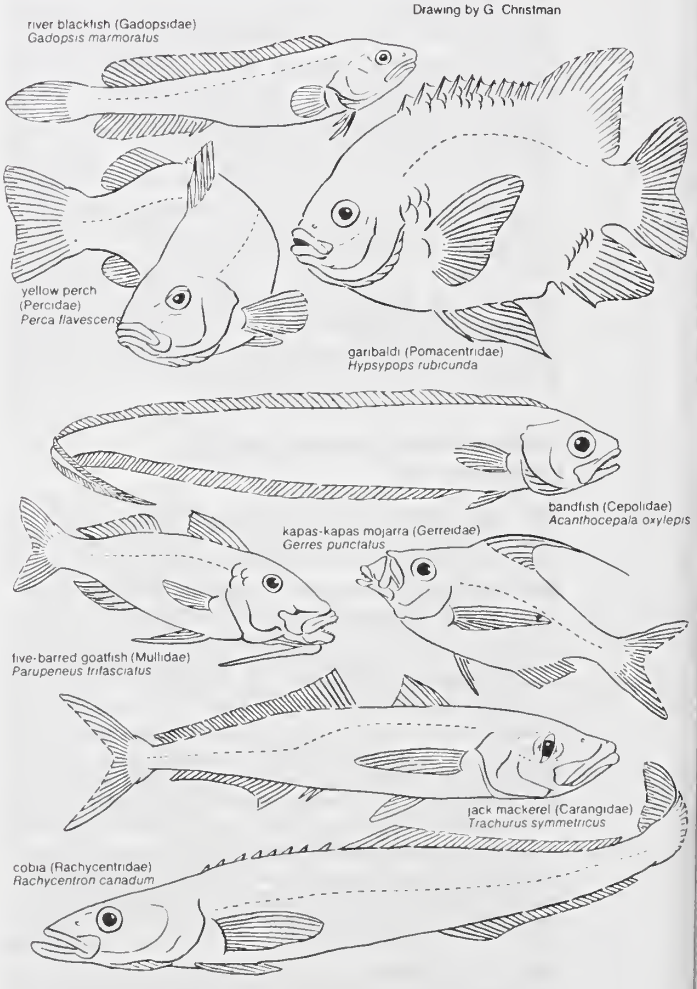


Figure 35: Representative perciforms of the families Pomacentridae, Percidae, Cepolidae, Mullidae, Gerreidae, Carangidae, Rachycentridae, and Gadopsidae.

blennies) and has brighter coloration of the fins and body. Black, white, green, red, blue, and silver are colours characteristic of the brightly coloured males of damselfishes (Pomacentridae), wrasses (Labridae), labyrinth fishes, and cichlids. Structural differences between the sexes vary from easily observed characteristics, such as the presence of a longer dorsal fin in the male dragonets, to less obvious characteristics, such as larger canines in the dentition of male blennies (Blenniidae) and gobies. Bands, blotches, and tinged markings may also be characteristic of the brighter coloured males, as in some wrasses. During the spawning season, males of certain species of cichlids, parrotfishes (Scaridae), sparids, and wrasses develop a swelling on the forehead that may persist throughout life. These characteristics presumably enhance the male in its breeding and courtship activities.

Little is known of the courtship activities in most marine species of perciforms. Among those that have been studied, the male dragonet, with extended fins and gill covers, performs a display of colours while swimming repeatedly around the female. A similar courtship activity is also carried out by the fighting fishes (*Betta*).

Fertilization is usually external, although internal fertilization, in which the eggs are fertilized within the body of the female, is well known in such groups as surfperches (Embiotocidae); and the male in such cases generally possesses an intromittent organ, which functions in the transfer of sperm to the female. Internal fertilization also occurs among some of the gobies, clinid blennies (Clinidae), and apogonids. Perciforms that undergo internal fertilization mostly are viviparous; *i.e.,* they give birth to live young. The number of young in viviparous perciforms varies from three to 50 in the surfperches. The larger viviparous forms, however, have been found to produce a greater number of young.

Nest-building behaviour

Most perciform fishes are oviparous—*i.e.,* they lay eggs that are fertilized externally. The number of eggs laid varies from a few hundred to more than 3,000,000 in a 32-pound (15-kilogram) yellowtail (*Seriola dorsalis,* Carangidae). Often, the eggs are released to float freely, but many species have evolved elaborate nest-building behaviour. Nest construction frequently consists merely of clearing away of a small area under rocks, which may be on the open bottom or even inside empty animal shells. The sunfishes and darters (Percidae) use their fins or body to dig a circular depression for use as a nest. Wrasses construct a nest out of stones, shells, and seaweed. Males generally undertake the task of building the nest, but in many cases both male and female share the labour.

Most labyrinth fishes build bubble nests, the procedure being similar among members of the family. The male Siamese fighting fish takes a bubble of air into his mouth, coats it with a mucous secretion, then blows the coated bubble to the surface; this process is repeated until a bubble nest is formed. After pairing, the female allows the fertilized eggs to drop to the bottom, where the male picks them up in his mouth and blows them into the bubble nest. Many marine perciforms produce pelagic eggs (*i.e.,* that float on or near the surface of the open sea). Almost all of the freshwater perciforms produce demersal eggs (*i.e.,* that sink to the bottom). A certain amount of adhesiveness in the demersal eggs keeps the eggs together in clusters; the elongated shape of the clusters of perch eggs helps in securing the clusters to aquatic plants and rock bottoms. Not all eggs become attached to aquatic plants and other objects; the mature male humphead (Kurtidae) possesses a hooked structure on his forehead, to which the cluster of eggs is attached as soon as the female produces them. Oral incubation, in which the eggs are held in the mouth of one of the parents, is found in certain species of cardinal fishes (Apogonidae), jawfishes (Opisthognathidae), labyrinth fishes, and cichlids. The male, female, or both may incubate the eggs orally until they hatch, after which the young may be mouthbreeders. Similarly, guarding of the nest sites may be undertaken by the male, female, or both parents; however, males of some sunfishes, darters (*Etheostoma*), and the Siamese fighting fish defend their nest against intruders. In addition to guarding the nest, certain perciforms also aerate the eggs by directing a flow of water into the nest with fanning movements of the fins.

Although there is no evidence of parental care in perciforms that produce pelagic eggs, a strong protective behaviour is shown by most perciforms that build nests or carry their eggs around with them. The male dwarf cichlid may help in the care of the young, but it is the female that looks after the eggs, removing dead eggs from the clutch. In certain other cichlids (*Apistogramma* species, for example) the female may help free the young from the eggs by gently chewing off the egg shells. The young of many cichlids follow their mother around and quickly enter her mouth should danger threaten. The discus and the freshwater angelfishes of the cichlid family feed their young on mucous secretions of their own bodies. The male Betta guards the young until they can swim away freely on their own.

Threat displays of gobies and blennies

Behaviour. *Territorial activity.* Territorial behaviour is found in many perciforms, especially during the breeding season, when the male, and in some cases the female, displays territorial behaviour in guarding the nest of eggs or the young; such fishes include certain cichlids, sunfishes, and darters. The young tigerfish (Theraponidae) protects a restricted area around a small hole dug by using its body; such territorial behaviour disappears when the tigerfish grows beyond a length of about nine centimetres ($3\frac{1}{2}$ inches). An intruder approaching the burrow of a jawfish is usually greeted by a threatening pose of flared gill covers and erected fins. Gobies and blennies are also known for their marked territorial display; peck order may be present among gobies holding territories, with the highest degree of competition between male gobies of the same size. The characteristic threats of gobies and blennies include flaring gill covers, gaping jaws, puffing of throats, head raising, and shaking of bodies. When threat displays

Drawing by G. Christman

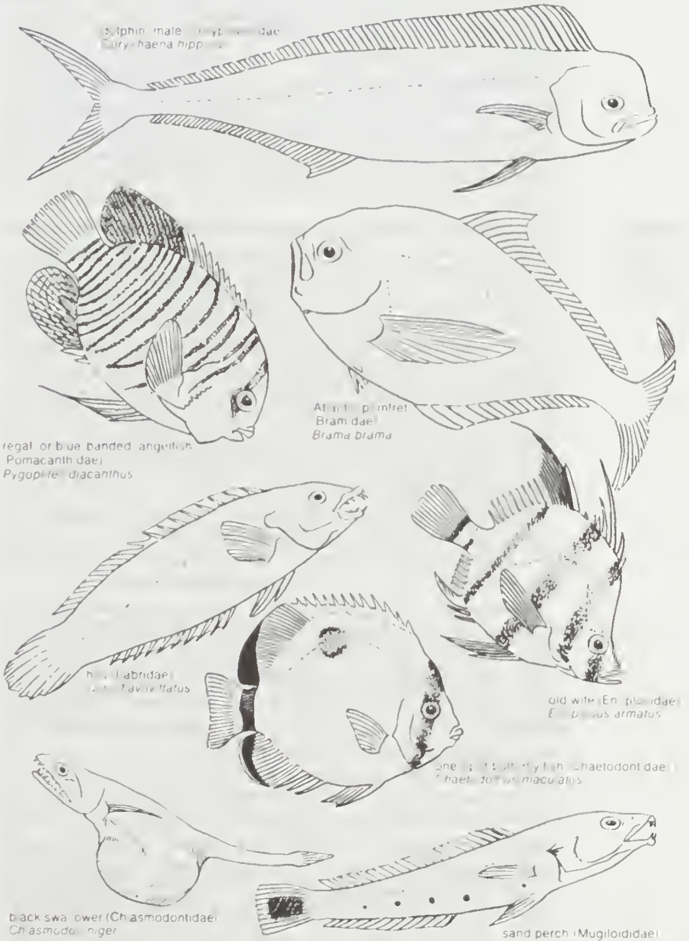


Figure 36: Representative perciforms of the families Coryphaenidae, Bramidae, Pomacanthidae, Enoplosidae, Chaetodontidae, Labridae, Mugiloididae, and Chiasmodontidae.

fail to settle a territorial dispute, male gobies fight, biting and chasing each other.

Sound production. The significance of sound production among perciform fishes is not well known, but most acoustic activity seems to be related to feeding and spawning periods. The level of sound production in croakers (Sciaenidae) increases considerably in the spawning season during the hours of late evening. There is also a difference in level between day and night; this may result from their feeding time. Damselfishes produce clicking sounds during feeding time, grinding the pharyngeal teeth. Another type of sound produced during feeding can be heard when parrotfishes feed on plant material covering reefs, biting on coral with their powerful platelike teeth. Grunts produce sounds by grinding their upper and lower pharyngeal teeth; the sounds are in turn amplified by the air (swim) bladder. Croakers, however, produce sounds by vibrating muscles of the abdomen that are attached to the sides of the air bladder, amplifying the vibrations of other muscles. The tigerfishes, or grunters (Theraponidae), have a similar system for sound production.

Feeding behaviour. Perciforms include both predator and prey species and are thus of great importance within the ecological food chains. The diverse adaptations for feeding are partly responsible for the success of this abundant order. Many of the colourful perciforms that occur around coral reefs are herbivorous fishes, the food of which consists mainly of plankton, algae on corals, and other reef vegetation; such fishes include parrotfishes, damselfishes, butterfly fishes, rabbitfishes and surgeonfishes. Among freshwater perciforms, certain species of *Tilapia* depend on aquatic plants for food. Most freshwater perciforms, however, are carnivorous, taking mosquitoes, insect larvae, and small insects. The larger predatory perciforms, in both freshwater and saltwater, feed on smaller fishes and even on birds and small mammals. They occupy a higher position within the food chain; examples include barracudas, groupers, tunas, and billfishes. The dolphins (*Coryphaena*) use their speed to catch fast prey such as flying fishes (Exocoetidae). The bluefish (*Pomatomus saltatrix*, Pomatomidae) is known for its voracious feeding behaviour; it feeds on open-water schooling fishes and, for unknown reasons, will continue to kill food fishes after its hunger is satiated.

Feeding behaviour of archer fishes

An interesting means of securing food is seen in archer fish (*Toxotes*, Toxotidae). The structure of the mouth in the archer fish is modified to form a groove along the roof of the mouth, against which the tongue fits to form a tube. The fish is able to direct a drop of water with remarkable accuracy at insects clinging to vegetation above the water surface. Thus bombarded, the insects fall into the water where they are quickly seized. Similar but less powerful squirting behaviour is also found in the butterfly fish. Another interesting type of feeding behaviour is seen in an African cichlid, which practices lepidophagy, the eating of scales plucked from other fishes.

Some predators lie in wait for their prey instead of pursuing it. An outgrowth of the mouth of the stargazer (*Uranoscopus scaber*) acts as a lure for prey. Groupers are also known to lie in wait for prey among rocks.

Adaptations of the mouth and jaw structure are seen in many of the perciforms. The piscivorous nandids (Nandidae) and the leaf fishes (Polycentridae) have large protrusible mouths capable of taking prey two-thirds their size, and the deeply cleft mouth of the swallowers (Chiasmodontidae) permits them to pass prey larger than themselves into their highly distensible stomachs.

Interspecific relationships. Mutual relationships among species are found in many perciform fishes. The cleaner fishes of the wrasse genus *Labroides* (Labridae) are well-known for their role in the removal of parasites from larger carnivorous fishes. The larger fishes recognize the cleaner fish and will not devour it. They allow free passage into their cavernous mouths and gill chambers, in which the cleaner fish feeds upon leftovers and parasites. Each *Labroides* maintains a "cleaning station," which is visited regularly by larger fishes such as groupers, eels, jacks, and snappers. A relationship of a protective nature exists in the butterfishes (Stromateidae), the young of which are often found among the tentacles of jellyfishes; the fishes are immune to the stings of the jellyfishes. Fry of horse mackerel and tuna (Scombridae) have been also found among the tentacles of jellyfishes. A similar relationship is seen in the clown anemone fish (*Amphiprion percula*), which is found among the tentacles of sea anemones. The mucous substances secreted by the anemone fish protect it from the stinging cells of the sea anemone. Some anemone fishes seek out only one type of sea anemone; others do not show any species preference. The sleepers of the genus *Vireosa* (Eleotridae) are usually found close to rock oysters and clams, into which they quickly disappear when danger threatens. A similar relationship exists between certain sea cucumbers (sac-shaped echinoderms of the class Holothuroidea) and cucumber fishes (Carapidae). These fishes are found among starfishes, clams, and sea urchins, as well as sea cucumbers. Some are host specific and may even parasitize the host, as in the Florida cucumber fish (*Carapus bermudensis*), which seeks out a specific sea cucumber of the genus *Actinopyga*, within which the cucumber fish makes its home. At times *Carapus* also feeds on the internal organs of the sea cucumber; this does not really harm the host because it regenerates the lost parts.

The blind goby, *Typhlogobius californiensis*, depends entirely upon holes dug by the ghost shrimp (*Callianassa*) for a home, and is unable to live without its help. Other gobies are known to share holes with burrowing worms, pea crabs, and snapping shrimps.

Imitative resemblance among perciforms

Certain perciform fishes depend upon imitative resemblance for survival. Immature tripletails (Lobotidae) will turn on their sides and float on the surface of the water, resembling dead leaves; similar behaviour is found in the leaf fish *Monocirrhus polycanthus* (Nandidae). Some wrasses (Labridae) resemble green algae because of their body coloration, a mixture of white, green, and brown. A remarkable mimic is seen in the case of the sabre-toothed blenny (*Aspidontus taeniatus*), which mimics the cleaner fish *Labroides*. By resembling a cleaner fish, the blenny is able to approach other fishes and surprise them by rushing in to bite off a piece of fin (see MIMICRY). Similar mimicry also occurs in an East Indies species of blenny that mimics a wrasse, apparently for food and protection.

FORM AND FUNCTION

The nature and diversity of the perciforms make a general definition of the group difficult; the most common characters are found in the large families of sea basses, mackerels, perches, sunfishes, and others. Perciform fishes usually have spines present on their dorsal, anal, and pelvic fins. The dorsal fin is usually divided into two parts, with the first part supported by one or more spiny rays; these are believed to have evolved for defense purposes. The pelvic fins are usually present, directly below or a little ahead of the pectoral fins, and they are supported by one spine and five or fewer soft rays. This position of the pelvic fins gives the perciforms an advantage in manoeuvring over short distances. The pelvic fins are lacking in some perciforms; in others, such as gobies, they are united to form a cuplike sucker; and, in the gouramis, the pelvic fin may be drawn out into long filaments.

A diversity of mouth and jaw structure occurs in the perciforms; most of it is brought about by the various types of feeding behaviour. Perciforms usually have protrusible jaws; and in the leaf fishes and swallowers, the jaws are easily distensible. The protrusible jaw may have thick lips, as in the wrasses, or may possess fleshy projections, as in certain species of African cichlids. Weever fishes (*Trachinus*) and stargazers (*Uranoscopus*) possess jaws that are directed upward; the jaws help when capturing prey as they lie buried in the sand. The upper jaws are greatly prolonged in the swordfishes and billfishes; the significance of this feature is rather uncertain. Many of the perciform species that inhabit coral reefs have modifications of the snout and jaws; the butterfly fishes have a straight tubelike mouth for reaching food among coral crevices.

Other structures of the perciforms have also undergone modification according to the various types of feeding behaviour. Most of the piscivores possess numerous short, fine, and pointed teeth; *e.g.*, the perches and sea basses.

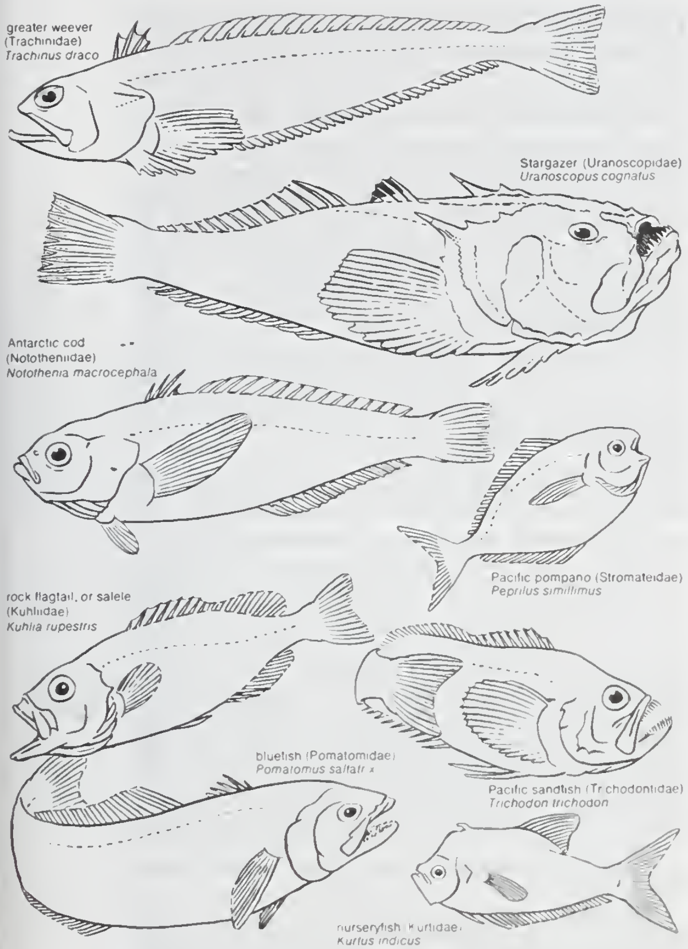


Figure 37: Representative perciforms of the families Kuhliidae, Pomatomidae, Trichodontidae, Trachinidae, Uranoscopidae, Nototheniidae, Stromateidae, and Kurtidae.

Drawing by G. Christman

Barracudas have long piercing canine teeth for holding and stabbing prey, and certain gobies and blennies characteristically have long, curved canines found in the lower jaw only. Perciforms that are either herbivorous or consumers of small invertebrates in addition to vegetation possess incisors, which are chisel-like teeth, as in certain sea breams; incisors may become fused into a beaklike structure, as in the parrotfishes. Enlarged pharyngeal (throat) teeth are present in some species of perciforms and are used for grinding and crushing hard-shelled food such as clams and snails. Tooth structure has undergone various modifications in many of the African cichlids. Herbivorous cichlids possess chisel-like teeth that are used to feed on plants and algae; the piscivorous ones have strong, pointed teeth. Cichlids that feed on the eggs and young of other species possess a highly distensible mouth with reduced teeth embedded in the gums. (M.W.L.)

CLASSIFICATION

Annotated classification. The classification presented here is mainly that of British ichthyologist P.H. Greenwood and colleagues, published in 1966; it in turn is fundamentally that established by the British ichthyologist C. Tate Regan in 1929. Certain changes have been incorporated, based mostly on characters from the nervous system.

All of the features distinguishing the order Perciformes are found in the fishes of the most generalized group, the suborder Percoidei, which contains the sea basses, sunfishes, perches, and fishes of many other families. As the subordinal name implies, the fishes composing it are "percoid," or perchlike in appearance. The fishes in the other suborders have presumably evolved from a percoid-like ancestor, but some have changed so much as to hardly resemble a percoid fish externally.

ORDER PERCIFORMES

Swim bladder not connected by an open duct to the throat; dorsal, anal, and pelvic fins usually with spines; dorsal fin usually with the first or anterior part supported by spiny rays and the rest by soft (articulated) rays, the spinous- and soft-rayed portions often separated from each other so as to constitute two distinct dorsal fins, or there may be a notch in the profile of the dorsal fin that indicates the two joined portions. A fin is considered long- or short-based on the basis of the length of its attachment to the body. Pelvic fins with 1 spine and 5 or fewer soft rays, or pelvic fins absent; pelvic fins thoracic in position (*i.e.*, placed ventrally below the base of the pectoral fins), pelvic fins sometimes ahead of pectoral fins (that is, jugular in position), or, occasionally, pelvic fins posterior to pectoral fins; pelvic girdle usually directly attached to base of pectoral girdle; caudal fin with not more than 17 principal (branched) rays supporting it; skull lacking orbitosphenoid bone; shoulder girdle lacking mesocoracoid bone; jaws typically protrusible; premaxillary bone of upper jaw excluding the maxilla from the gape of the mouth. Scales usually rough-edged (ctenoid), provided with small teeth (ctenii) along their posterior edge, sometimes round and smooth (cycloid). About 6,000 species, marine and freshwater; most species along shorelines in tropics and temperate zones, and in freshwater, the number of species dropping off drastically in higher latitudes. Fossil remains from the Upper Cretaceous Period (from 101,000,000 to 65,000,000 years ago).

Suborder Mugiloidei

Spiny-rayed dorsal fin rather widely separated from soft-rayed dorsal; pelvic fins of 1 spine and 5 rays, not thoracic but located more posteriorly on the abdomen; scaled lateral line with nerve pattern resembling that of some lower nonperciform fishes; taste nerves of trunk like many percoids; olfactory bulbs far forward, unlike other perciforms.

Family Mugilidae (mullets). Lower Oligocene to present; with a cigar-shaped, roundish body, short snout; large cycloid scales; usually numerous small movable teeth; muscular gizzard-like stomach in many species. Moderately large schooling fishes, 30 to 90 cm (1 to 3 ft) long. Less than 100 species; tropical and temperate waters, some in brackish water and some in fresh water.

Suborder Sphyraenoidei

Two dorsal fins, both short-based and widely separated from each other; pelvic fins some distance posterior to pectorals; pattern of trunk lateral line nerves in a rudimentary percoid pattern resembling pre-percoid patterns found in atheriniform fishes; pattern of taste nerves on trunk resembles that in atheriniform fishes; large teeth set in deep sockets.

Family Sphyraenidae (barracudas). Eocene to present; large, elongated, pikelike, with long, pointed jaws and big teeth; piscivorous; probably not over 120 cm (4 ft) long; all warm seas; about 20 species; fine game fishes.

Suborder Polynemoidei

Pelvic fins thoracic; pectorals low on side of body, divided into an upper normal part and an unusual lower part, consisting of a number of fin rays grown out into long sensory filaments, reaching to anal fin and far beyond in some species; pointed snout; large eyes.

Family Polynemidae (threadfins). Upper Miocene to present; resembling mullets in body shape and widely separated two dorsal fins, but like anchovies in ventral mouth with projecting snout, rather deeply cleft mouth, and adipose eyelids; in most warm seas, often abundant at river mouths and over sandy bottoms; about 24 species, most 30–69 cm (12–24 in.) long, but 1 giant species reaches 180 cm (6 ft).

Suborder Percoidei

The largest suborder both in numbers of families and in species; fishes typically of a perch or bass appearance; jaws protrusible; dorsal fin usually conspicuously spinous, often with the spinous and soft portions separated or nearly so or with a notch between them; anal fin with 2 or more spines at anterior end, occasionally with 1 spine, sometimes with more than 3 spines; pelvic with 1 spine and usually 5 soft rays; body often somewhat deep rather than elongated; territorial, bottom-oriented, investigative shorefishes with great close-quarters swimming manoeuvrability, swimming backward and forward short distances with much use of pectoral fins; great variety in adaptive design and operation of jaws; generalized predators of fishes and crustaceans in all warm seas near shores, especially in tropics; in freshwater. Almost 4,000 species, some of large size; about 90 families.

Superfamily Percoidea (basses, perches, sunfishes, cichlids, damselfishes, and many others).

About 60 families grouped together because they show no great morphological specialization away from the general bass, grouper, or perch kind of fish taken as a model. Most inhabit

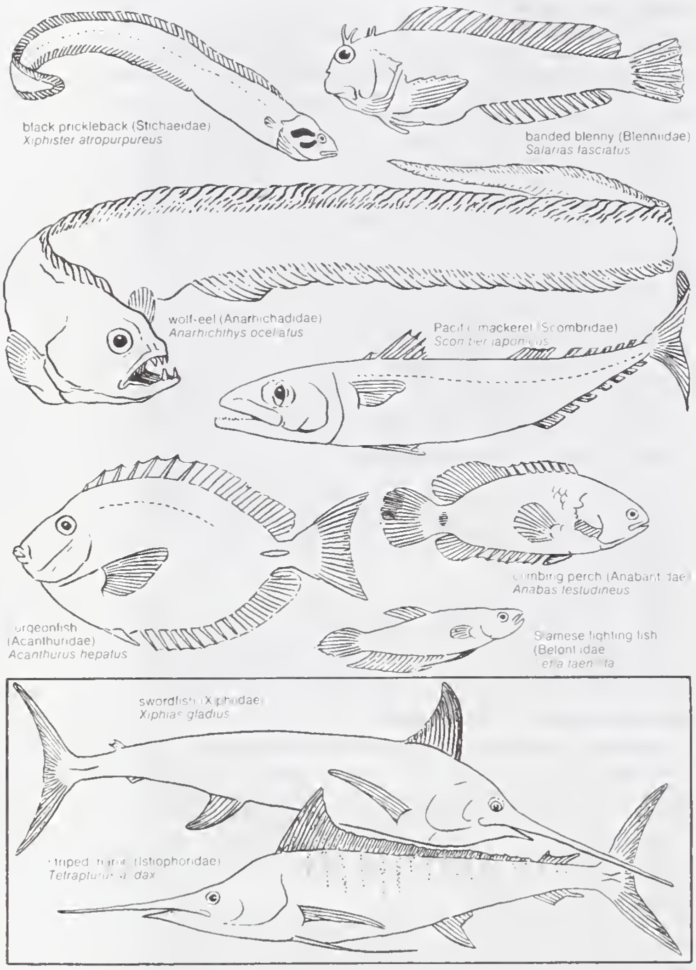


Figure 38: Representative perciforms of the families Blenniidae, Stichaeidae, Anarhichadidae, Acanthuridae, Scombridae, Xiphiidae, Istiophoridae, Anabantidae, and Belontiidae.
Drawing by G. Christman

shores of tropical and temperate seas or lakes. Two aspects most obviously contributing to the success of this group are the adaptability of the protractile mouth and the detailed variety of specializations for swimming manoeuvrability in restricted areas.

Family Scombropidae. Pliocene to present; rare, deep-water marine (down to 600–800 m, or 2,000–2,600 ft); this and the next several families retain some features that may have been those of the most generalized ancestors of present-day percoids such as: 2 dorsal fins separate, anal fin with 2 spines, primitive pattern of taste nerves on the body and remains of a special system of lateral line nerves on the head. About 7 genera and 20 species.

Family Acropomatidae. Rare, deep-water marine species similar to scombropids; anus located anteriorly from normal position at front of anal fin. Light organs present; midwater depths of 300–500 m (1,000–1,650 ft); 2 species; Indo-Pacific.

Family Apogonidae (cardinal fishes). Eocene to present; 2 well-separated dorsal fins; 2 spines in anal fins; midsection of body often short, with head, eyes, and caudal peduncle proportionately larger; mouth large. Males orally incubate eggs. Mostly marine species, often reddish; live around coral reefs in tropics and subtropics; nocturnal; a few species from deep oceanic midwaters 1,000–1,200 m (3,300–4,000 ft); most with special lateral line system of nerves and free organs well developed on head (among the few percoids with the system). About 200 species.

Family Moronidae (temperate basses). Eocene to present. Two dorsal fins connected at their bases. Most species slim-looking basses; well-known food and game fishes such as striped bass and white basses of the genus *Morone.* Some species anadromous. Weight to 50 kg (about 110 lb) in striped bass. About 12 species, marine and freshwaters of North America, Europe, Africa, Australia, and the Orient.

Families Pomadasyidae and Banjosidae (grunts). Eocene to present. Spinous and soft dorsal fins continuous, often notched. Resemble snappers (Lutjanidae), but teeth weaker, canines absent; sound produced by grinding pharyngeal teeth and amplified by adjacent swim bladder. Second anal spine often much enlarged; about 75 species, tropical and subtropical shorefishes, many entering estuaries; good food and game fishes.

Family Centropomidae (snooks or robalos). Eocene to present. Elongated, basslike fishes; head long and sloping; sizable mouth, projecting lower jaw; 2 separate short-based dorsal fins; short-based anal fin. Oceans and estuaries of Pacific and Atlantic coasts of tropical Americas; about 8 species, 45–150 cm (about 1 ½ to 5 ft); good food and game species.

Family Dinolestiidae. One species, resembling, but not related to, the barracudas (Sphyraenidae, above). Marine; Australia and Tasmania; length to 50 cm (20 in.).

Family Latidae (Nile perches). Eocene to present. Closely similar to Centropomidae. Piscivorous; freshwaters of Africa; at river mouths and along coasts in Southeast Asia and northern Australia; up to about 180 cm (6 ft) long and 140 kg (300 lb); excellent food fishes.

Family Percichthyidae (perch trouts). Eocene to present. Dull-coloured, small, perchlike freshwater and marine fishes of Chile and Argentina. Dorsal fin deeply notched. About 9 species.

Family Ambassidae (glass perches). Small fishes similar in body form to Apogonidae; body short, rather deep. Spinous and soft-rayed parts of dorsal fin nearly separated by deep notch. Good predators on mosquitoes; large schools in sea around river mouths and in freshwater; Indo-Pacific; about 24 species.

Family Serranidae (sea basses, groupers, and many others). Paleocene to present. Variously similar to many of the percoid families mentioned above in their general spiny-rayed, perchlike appearance. Dorsal fin continuous, but may be deeply notched; spinous portion of dorsal fin with longer base than soft dorsal portion; anal fin usually with 3 spines and short-based; no scaly sheath along base of dorsal and anal fins; mouth large; pectorals broadly rounded; caudal fin usually truncate or rounded, sometimes moderately forked. About 400 species in tropical, subtropical, and warm temperate seas; some in freshwater; good food and game fishes; maximum weight to about 320 kg (700 lb).

Families Pseudochromidae, Grammidae, Plesiopidae, Pseudoplesiopidae, and Acanthoclinidae. Quite similar, small, darkly colourful, rather secretive coral-reef basslike fishes of tropical Indo-Pacific and Caribbean seas. An interesting specialization of numerous species is the presence of multiple horizontal, interrupted lateral lines on trunk: 1 along the back, 1 along the side, and 1 along the bottom of each side of the body. Dorsal and anal fins vary from few or no spines up to 24; long dorsal fin, sometimes deeply notched between spines with a little banner or flag of fin membrane extending up and out from the end of the spine. Together, about 40 species.

Families Glaucosomidae (pearl perches) and *Lobotidae* (triple tails). Deep-bodied perchlike fishes found in eastern Pacific Ocean, except *Lobotes,* which also occurs elsewhere in tropical salt and freshwater; 5 or 6 species in Glaucosomidae; 1 in Lobotidae.

Family Priacanthidae (big eyes or catalufas). Eocene to present. Deep-bodied, reddish, serranid-like dwellers of deeper offshore waters, toward the bottom. Jaws almost vertically hinged; carnivorous. Tropical Indo-Pacific and Atlantic; few species; 30–45 cm (12–18 in.) long.

Family Centrarchidae (sunfishes and basses). Eocene to present. Moderately deep-bodied; spinous and soft dorsal fins continuous, not separate as in Percidae; more than 3 anal spines. Freshwaters of North America; only 1 species, *Archoplites interruptus,* native west of the Rocky Mountains; various species widely introduced elsewhere; prefer quieter waters, such as ponds, lakes, swamps; excellent game fishes; size from 30 gm (1 oz) to about 10 kg (22 lb); 30 species.

Family Embiotocidae (surfperches). Miocene to present. Laterally compressed, ovate, smooth-scaled, fairly small, with small head. One long dorsal fin, depressible into a scaled sheath alongside the fin. Give birth to actively swimming young; 23 species from central Baja California to Japan, absent from Aleutian chain; most species (20) occur in California; 1 freshwater species (*Hysterocarpus traski*) in central California; 12–30 cm (5–12 in.).

Families Nandidae and Polycentridae (near leaf fishes and leaf fishes). Small, mostly piscivorous fishes with large to huge protrusible mouths; consume prey up to $^2/_3$ their own length. Bodies moderate to deep, laterally compressed; long spinous dorsal fin and 3 to 13 spines in anal fin; soft dorsal and anal fins short-based and colourless, used together with colourless pectoral fin in *Polycentrus* and *Monocirrhus* in swimming imperceptibly toward prey without any evident signs of fin or body movement. Six species in freshwaters (and brackish for *Nandus*) from India to Malaysia; for Polycentridae, West Africa and the northeast coast of South America. Six species.

Family Cichlidae (cichlids). Eocene to present. Small freshwater (a few brackish water) percoids, resembling damselfishes (below) and North American sunfishes. One pair of nostrils instead of the 2 of most fishes; spinous dorsal fin long-based; anal fin short-based, 3-spined; dorsal and anal fins often pointed at posterior ends; caudal fin usually rounded; first pelvic rays of males often elongated; lateral line interrupted. Native to Texas, Central America, South America, and West Indies; Africa; Palestine, India, and Sri Lanka (formerly Ceylon). Complicated courtship, nest-building and mouthbrooding (several genera). Over 700 species, many of them important aquarium fishes; almost 600 native to rift lakes in Africa (especially Lakes Tanganyika, Nyasa, and Victoria); a few species up to 30 cm (12 in.); 1 up to 9 kg (20 lb; genus *Tilapia*).

Family Pristolepidae. A deep-bodied, laterally compressed, small-mouthed percoid, resembling a cichlid or pomacentrid but having a patch of blunt, molariform teeth on base of skull opposite similar teeth on rear floor of mouth. Long spinous dorsal fin; anal fin 3-spined; interrupted lateral line. One species only, of unsolved percoid relationship but evidently not allied to Nandidae as once thought. Freshwaters; Burma to Indochina and Malaysia.

Family Pomacentridae (damselfishes and anemonefishes). Eocene to present. Abundant, conspicuous, active little fishes, often brightly coloured, found near shores and coral reefs, mainly in tropical but a few in subtropical seas. Resemble cichlids in general appearance and, like cichlids, have only 1 pair of nostrils. Lateral line continuous or interrupted; long spinous dorsal fin; anal fin with 2 spines, sometimes 3; soft dorsal and anal fins similar; caudal fin usually forked; all fins may have pointed filamentous ends; head with scales; scales large, ctenoid; bases of unpaired fins scaled; first ray of pelvic fin somewhat elongated; floor of mouth with 1 triangular, fused tooth plate in pharynx (pharyngeal plate). Territorial and pugnacious; about 250 species.

Family Percidae (perches, walleyes, darters). Eocene to present. Spinous and soft dorsal fins usually well separated; anal fin with 1 or 2 spines and short-based; scales ctenoid; bodies rather elongated. All freshwater, temperate species; perches and pike perches Holarctic with a few brackish-water species and a marine species of pike perch in parts of Black and Caspian seas; darters are native only to North America. Perches prefer quiet waters, darters running waters; pike perches occur in either and are semimigratory. Many species build nests and show parental care; size up to about 90 cm (3 ft) and 11 kg (25 lb) for walleyes; darters from 2.5 to 10 cm (1 to 4 in.). About 125 species, of which about 100 are darters.

Drawing by G. Christman

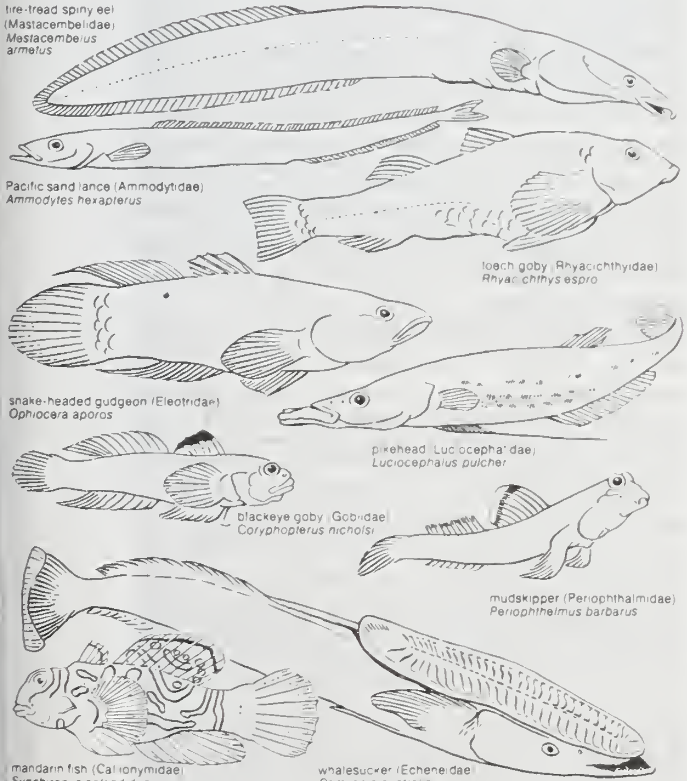


Figure 39: Representative perciforms of the families Luciocephalidae, Mastacembelidae, Echeneidae, Ammodytidae, Eleotridae, Gobiidae, Periophthalmidae, Rhyacichthyidae, and Callionymidae.

Family Sciaenidae (drums or croakers). Upper Cretaceous to present. Some species resemble cods, others resemble sea basses; most have lower jaw short or underslung, with upper jaw and snout extending beyond lower jaw; often 1 or more barbels (fleshy filaments) at tip of lower jaw. Spinous and soft dorsal fins separate; soft dorsal fin fairly long-based; anal fin small, with 2 spines (most percoids have 3 or more); lateral line continues out to posterior end of caudal fin (unusual for most percoids); air bladder often with intricate outpocketings and with muscles attached to it that operate to make resonating sounds in air bladder, hence name of drum; surface of head may be cavernous through expansion of lateral line canal system. Most species occur on slopes of continental shelf, a few around islands; most in tropics, a few in temperate waters; a few freshwater. About 160 species; size from about 100 gm to 100 kg (a few ounces to 220 lb); many are important food fishes.

Family Odacidae (rock whitings). Teeth incompletely fused together making a parrot-like beak as in parrotfishes (Scaridae), but odacids are evidently not related to wrasses (Labridae), parrotfishes, or whitings (Sillaginidae), all of which they also resemble. Seven species; southern Australian and New Zealand seas.

Family Labracoglossidae. Resemble grunts (Pomadasyidae), to which they are allied. Five species. Easter Island to New Zealand, Australia, Japan.

Family Sillaginidae (whitings). Oligocene to present. Elongated fishes with long, conical snout, small mouth; moderately long dorsal and anal fins; anal fin with 2 weak spines. About 6 species of small marine fishes of shallow water; Indo-Pacific, often in estuaries and river mouths; dig in bottom with long snouts for food.

Family Branchiostegidae (tilefishes). Pliocene to present. Body elongated; large, oblique mouth with strong canines; body deep through chest region; eyes high on head at top of steep sloping forehead; single, rather long dorsal fin; fin spines weak. Moderate to large body size; about 20 species, most in shallow seas of tropics and temperate zone.

Family Lactariidae (milk trevally). Miocene to present. Moderately deep-bodied, laterally compressed; mouth large, oblique; eyes large; pectorals pointed; 2 dorsal fins separated; anal fin long-based. One or 2 species marine in Indo-Pacific.

Families Owstoniidae and Cepolidae (bandfishes). Eocene to present. Owstoniids are marine, deep-water fishes, basslike, but large mouth is oblique, eyes large, and dorsal and anal fins long, continuous, and high; caudal fin with long rays; body tapers noticeably. Cepolids similar, but have a long tapering body and are shallow- and deep-water fishes. Cepolids occur from Europe through Mediterranean to India, China, Japan, and Philippines; owstoniids only in Far East; few species.

Family Mullidae (goatfishes). Miocene to present. Resemble minnows (Cyprinidae); have a long pair of chin barbels that usually lie flat against chin, except when in use as sense organs, probing the bottom for food. Spinous dorsal fin well separated from soft dorsal fin. Fifty to 60 species at reefs and shallow sandy or muddy bottoms near shore in tropics and warmer temperate seas.

Family Lutjanidae (snappers). Miocene to present. Resemble sea basses (Serranidae), but when mouth is closed jaw slips under bony cover over preorbital area (between eyes and jaw); enlarged canine teeth in jaws; spinous dorsal fin longer than soft fin and joined to it; anal fin short-based; caudal fin usually truncate. About 250 species; marine and brackish water in all warm seas.

Families Nemipteridae, Scolopsidae, Lethrinidae, Pentapodidae. Resemble Lutjanidae; some with wider preorbital area under which upper jaw slips; others (Nemipteridae) with molar teeth in sides of jaws and incisors or canine-like teeth at front end of jaws. About 50 species collectively; marine, Atlantic and Indo-Pacific, especially around coral reefs; some species up to 90 cm (3 ft) long; good food fishes.

Family Sparidae (breams, porgies). Eocene to present; resembling fishes of families Nemipteridae and Lutjanidae; mouth small but with a powerful dentition of incisors or canines across the front of the jaws and molar-like teeth to the sides, the teeth enlarging toward the rear of the mouth; used in crushing crustaceans, mollusks, and small fishes. About 100 species in shallow seas of tropical, subtropical, and temperate zones; most species less than 30 cm (1 ft) long, a few up to 120 cm (4 ft); important food and game fishes.

Family Gerreidae (mojarras). Small perchlike fishes with compressed, rather deep, silvery bodies; they are set off from most other percoids by their highly protrusible mouths used in probing soft, sandy bottoms to catch food organisms; about 40–50 species, mostly tropical marine, a few in freshwater; abundant in American tropics and East Indian region. Resemble fishes of family Leiognathidae, but the similarity is thought to be the result of parallel evolution.

Family Pempheridae (sweepers). Perchlike fishes with compressed body and a very short-based dorsal fin but long-based, low anal fin; big eyes. About 20 species in tropics of Atlantic, Indian, and Pacific oceans.

Family Bathyclupeidae. Resemble sweepers but apparently not related; compressed body; prominent lower jaw; single short-based dorsal fin; anal fin long-based; eyes large; mouth large. Few species; deep-sea midwaters, 400–500 m (1,300 to 1,640 ft) depth.

Family Carangidae (jacks, scads, and pompanos). Eocene to present. Lateral line usually strongly arched anteriorly and the posterior part usually armed with enlarged, keeled, scutelike bony scales; second dorsal and anal fins elevated and falcate anteriorly and sometimes followed by short-based free finlets; soft-rayed portions of dorsal and anal fins preceded by separated or nearly separated free, short spines; swift-looking fishes with streamlined bodies, very constricted caudal peduncle, deeply forked caudal fin, and long, sickle-shaped pectoral fins; eyes often covered with clear eyelids; worldwide in tropics and warm temperate areas; over 200 species; most are good food and game fishes.

Family Rachycentridae (cobias). Long and slender, somewhat resembling mackerels and shark suckers (remoras); may be related to shark suckers (family Echeneidae); soft portion of dorsal fin preceded by a row of short, separate dorsal spines. One species, worldwide in warm seas but absent in eastern Pacific; size up to 90 cm (3 ft) and over 70 kg (154 lb); pelagic in habit.

Family Coryphaenidae (dolphins). Sleek, fast-swimming fishes of tropic and temperate open seas; speeds up to 65 km (40 mi) per hour; dorsal and anal fins long-based and low; dorsal fin begins on head; deeply forked caudal fin; long, slender pelvic fins fitting into groove on belly; young males of one species change into adults with almost vertical forehead. Two species, 1 large, up to 150 cm (5 ft) and 25 kg (55 lb).

Family Formionidae (black pomfrets). Body deep, strongly compressed; dorsal and anal fins long-based and elevated at anterior end, both fins with spines rudimentary in adult fish; a few enlarged scutes at end of lateral line, a characteristic of fishes of jack family (Carangidae). One species, up to about 60 cm (2 ft); Indian Ocean to Australia and China.

Family Menidae (moonfishes). Eocene to present. Body strongly compressed, triangular, very deep behind head, with edge of chest razor-sharp; anal fin very long and low, reaching forward to pelvic fins; first ray of pelvic fin long and filamentous; mouth protrusible into an upward extended tube; 1 species, less than 30 cm (12 in.) long, in Indo-Pacific seas.

Family Leiognathidae (slipmouths). Oligocene to present. Small fishes; body greatly compressed, ovate; mouth small but highly protrusible into a tube pointing up or down; long dorsal and anal fins elevated anteriorly, folding down into scaly sheath lying alongside bases of the fins; exude large amounts of mucus after capture. About 25 species, mostly marine shallow-water shorefishes, some in estuaries; tropical Indo-Pacific area.

Family Bramidae (pomfrets). Body deep, strongly compressed; head broadly rounded in profile; pectorals long and curved; long-based dorsal and anal fins, low, sloping back to narrow caudal peduncle and deeply forked caudal fin; mouth small, oblique; about 23 species of darkly coloured fishes in rather deep waters of open ocean; size up to 90 cm (3 ft). One species, *Pteraclis velifera,* with enormously high and long fanlike dorsal and anal fins.

Family Caristiidae (manefishes). Rare black pomfret-like fish from midwater depth of 1,000 m (3,300 ft) over much deeper bottoms; dorsal fin begins far forward over end of cranium, high and like a mane; pelvic fins very long; 1 or 2 oceanic species.

Family Monodactylidae (fingerfishes). Includes family Psettidae. Small, silvery, ovate, small-mouthed fishes of salt and brackish water, temporarily in freshwater. Pelvic fins minute; dorsal and anal fins long; anterior spines of dorsal fin short and detached. Five species, coasts of Africa to India and Southeast Asia.

Family Toxotidae (archerfishes). Lower Tertiary to present. Moderately deep-bodied percoids distinguished by nearly straight line from dorsal fin to tip of jaws; jaws large, oblique; lower jaw prominent; jaws and roof of mouth adapted for expelling drops of water fired from the surface at insects on overhanging tropical vegetation in mangrove swamps. Dorsal and anal fins set far back on body; dorsal fin with 4 enlarged spines at front; 5 species, with black saddle markings or spots on dorsal surface; marine and freshwaters of tropical Indo-Pacific area.

Family Ephippidae (spadefishes and batfishes). Includes Chaetodipteridae, Drepanidae, and Platacidae. Eocene to present. Body deep, orbicular, greatly compressed laterally; long-based dorsal and anal fins, usually high, especially anteriorly in young; mouth quite small. About 15 species in tropics of world, few in temperate zone; size up to 40 cm (16 in.).

Family Pomacanthidae (angelfishes). Eocene to present. Body deep, laterally compressed; mouth quite small; strongly resemble Pomacentridae and Chaetodontidae (in which family they traditionally have been placed) but distinguished from these and other deep-bodied coral-reef fishes by a rather long, sharp spine projecting posteriorly from the lower part of cheek region (from the preopercle bone); profile of head from dorsal fin to tip of snout mainly convex, rarely straight, not concave as in butterflyfishes; young often differently coloured from adults; less than 100 species of conspicuous coral-reef and tropical fishes; mostly of small size, a few species up to about 45 cm (18 in.).

Family Scatophagidae (scats). Eocene to present; deep-bodied, laterally compressed; small mouth; 2 dorsal fins joined at base; bases of dorsal and anal fins scaled; about 4 species; coasts of Africa to Indo-Pacific, entering brackish and freshwater; scavengers, eating decaying plant and animal remains and fecal matter, including human; size up to 40 cm (16 in.).

Family Rhinoprenidae. One species, recently discovered, superficially similar to Scatophagidae, but really quite different. Long free filament at front of dorsal and anal fins and 1 on pectoral fin; posterior nostril of each side of head enlarged and confluent across snout. One species over muddy bottoms at river mouths in Gulf of Papua, Papua New Guinea.

Family Enoplosidae (old wife). Eocene to present. Body laterally compressed; spinous and soft dorsal fins elevated anteriorly, as is anal fin; general appearance gives impression in side view of two separate bodies joined together at midpoint; pelvic fins large. One species in rocky areas of Australian coast; size up to 22 cm (9 in.).

Family Chaetodontidae (butterflyfishes). Oligocene to present. Body deep, disk-shaped; mouth small, jaws sometimes on end of small (or, occasionally, fairly long) beak; slope of forehead from dorsal fin to tip of snout is concave, never convex; spinous portion of dorsal and anal fins usually prominent; bases of dorsal and anal fins scaled. Body often marked with oblique dark bands and one or two eye-like spots. Over 100 species in tropics of world, all marine; typical of coral reefs.

Family Pentacerotidae (pelagic armorheads). Resemble Chaetodontidae but with higher dorsal fin and much larger dorsal, anal, and pelvic spines; dorsal fin usually strongly elevated; head rough with exposed bony plates; snout usually elongated bearing small mouth at its end; lips with "hairy" skin. Seven or 8 species, in deeper coastal waters from South Africa, Australia, and New Zealand to Japan; size up to about 50 cm (20 in.).

Superfamily Kuhlioidea

Classified together under this superfamily are a number of percoid families that have in common the same pattern of main branches of a complex of taste nerves that occur on the body.

Family Kuhliidae (flagtails and aholeholes). Typical percoids resembling North American freshwater sunfishes and basses; dorsal fin single, notch between soft and spinous portions; dorsal and anal fins folding into scaly sheaths along their bases; 6 species of tropical Indo-Pacific oceans, preferring brackish water and freshwater; size up to 45 cm (18 in.); good food fishes.

Families Scorpididae (sweeps and halfmoons), *Kyphosidae* (sea chubs), *Girellidae* (opaleyes), *Coracinidae* (galjoen fishes). All similar families recognized by combination of ovate body, small mouth, strong caudal fin that is usually weakly forked; and, especially, a spinous dorsal fin with low spines followed by a higher evenly curved or falcate soft dorsal fin; about 36 species, many partly herbivorous; tropical and warm temperate coasts in Atlantic, Pacific, Indian oceans; size usually not over 45 cm; some are good food and game fishes.

Family Oplegnathidae. Pliocene to present. Strongly resemble Scorpidae and Kyphosidae, but incisiform teeth of young become fused in adult to form a parrot-like beak to upper and lower jaws; these fishes are not related to true parrotfishes (Scaridae); several species of shorefishes mostly in tropics of Southern Hemisphere; size up to 60 cm (24 in.).

Family Theraponidae (grunters or tigerfishes). Typical percoids of small bass type; colours dull or silvery or with horizontal dark stripes; dorsal fin notched, spinous part longer than soft part; some species make grunting sounds. Less than 25 species, Indian and western Pacific oceans and in fresh waters of Australia and New Guinea; small to medium size.

Family Arripidae (Australian salmon; ruffs). Not related to true salmons of Northern Hemisphere. Rather long, slender body; deeply forked tail; moderately long dorsal fin, a notch between the shorter spinous dorsal and longer soft dorsal fin. Two species; marine, young in brackish water; shallow waters off South Australia, New Zealand, and adjacent islands; size up to 1 m (3 ft); important food and game fishes.

Family Leptobramidae (salmon trout). A slender carangid-like species with large mouth, rather long-based anal fin, and a single dorsal fin placed behind the beginning of the anal fin; resembles Pempheridae but apparently is not related to it; a single species reaching 43 cm (17 in.) and about 2 kg (4 lb); off coasts of western Australia and New Guinea.

Family Emmelichthyidae (bonnetmouths). Includes families Caesionidae, Erythricthyidae, Dipterygonotidae, Maenidae, Spicaridae, Centracanthidae, Merolepidae by some authors. About 25 species of 2 general body types: one with slender, elongated bodies with moderately protrusible upper jaws; the other deeper bodied and with enormously protrusible upper jaws. Some school in open waters at depths of 3–45 m (10–150 ft), others in deeper water.

Family Pomatomidae (bluefish). Resembles Australian salmon (family Arripidae), but spinous dorsal smaller and separate from soft dorsal; anal fin longer; body more robust; 1 or 2 small spines before anal fin. One species, widely distributed in tropical and warm temperate seas; voracious feeder on other fishes; size up to 120 cm and 11 kg (4 ft and 25 lb); good game and food fish.

Family Nematistiidae (roosterfish or papagallo). Streamlined fish resembling jacks (Carangidae); dorsal fin remarkable for spinous portion consisting of greatly lengthened spines nearly separate from each other, the fin connected by fin membrane only at their bases, except the last ray, which is free at its base. Caudal fin deeply forked; pectoral fin long, falcate; mouth rather large. One species found only on west coast of Central America, from Gulf of California south to Panama; frequents sandy shores; size up to 35 kg (80 lb); game fish.

Superfamily Labroidea

Family Labridae (wrasses). Paleocene to present. Various body forms, but commonly cigar-shaped, fairly slender; snout moderately long, sometimes lengthened; dorsal and anal fins long-based and low; caudal fin truncate; especially characteristic are the noticeable lips and outcurving sizable canine teeth at ends of upper and lower jaws, the large scales and the habit of swimming around the coral reefs by "rowing" with the pectoral fins. About 300 species, many beautifully coloured and marked, often with differences of colour and pattern between sexes and between young and adults; size from 5 cm to nearly 2 m (2 in. to several ft), size up to 10 kg (22 lb); marine shallow-water fishes of tropics and warm temperate zone.

Family Scaridae (parrotfishes). Eocene to present; resemble Labridae but with stouter bodies and with teeth fused to form a parrotlike beak to upper and lower jaws; large scales in regular rows; herbivorous; about 80 species, often brilliantly coloured; species separable by colour and pattern; sexes often differ in colour. Size up to 120 cm (4 ft) and 20 kg (45 lb); tropical marine fishes.

Superfamily Gadopsoidea

Family Gadopsidae (river cod). Slender, somewhat elongated, with longer based dorsal and anal fins than typical percoid. Characterized by very slender and fairly long pelvic fins located in front of pectoral fins; caudal fin rounded. One species, rivers of southern half of Australia; somewhat resembles Ophidiidae (order Paracanthopterygii) but apparently is a derivative of family Serranidae; size up to 60 cm (2 ft).

Superfamily Cirrhitoidea

Family Cirrhitidae (hawkfishes). Small colourful perchlike fishes having lower rays of pectoral fins unbranched, thick-ended, and separate from one another; small flag of skin projects from tip of each spine of dorsal fin; about 35 species; shallow coastal waters in warm seas.

Families Chironemidae, Haplodactylidae, Cheilodactylidae, and Latridae. Similar to Cirrhitidae; 25 to 30 marine species, mostly in shallow waters of Australia and New Zealand, some species off South America; size up to 60 cm (2 ft).

Superfamily Champsodontoidea

Family Champsodontidae. Small, elongated spiny-rayed fishes with a small spinous first dorsal fin and rather long, soft dorsal and anal fins; pelvic fins rather large; eyes near top of head and close together; unusually large mouth, the jaw extending obliquely past the eyes; several species; carnivorous; deep waters of Indo-Pacific oceans.

Superfamily Trichodontoidea

Family Trichodontidae (sandfishes). Resemble codfishes, but eyes high on side of head; mouth large, oblique; lips fringed; pectoral fins with long base extending forward past pelvic fins. Two species; marine; North Pacific; to 25 cm (10 in.).

Superfamily Trachinoidea

Family Opistognathidae (jawfishes). Resemble Clinidae, but jaws large to huge, extending far past eye; dorsal fin long-based; spinous and soft portions continuous; anal fin long-based; body usually elongated, slender; eyes almost at anterior tip of head; pelvic fins below pectorals. About 24 species, mostly small, in shallow tropical and temperate seas.

Family Bathymasteridae (ronquils). Resemble Opistognathidae, but jaws not so large; no spines in dorsal or anal fins; pelvic fins slightly ahead of pectorals; about 8 species; bottom-dwelling; coasts of North Pacific Ocean.

Family Mugiloididae. Includes Parapercidae and Pinguipedidae. Some resemble labrids in long dorsal and anal fins (sometimes with few spines), enlarged lips that appear to curl back, and enlarged canines at front of jaws. Body elongated, cylindrical; usually spotted and banded; eyes near top of head. Size from small up to 60–90 cm (2–3 ft); about 30 species; marine; bottom dwellers, coasts of South America, South Africa, Indo-Pacific to Japan; a few good food species.

Family Percophiidae. Includes Bembropidae and Hemerocoetidae, resemble flathcads (Platycephalidae); body long, slender; head flattened; eyes on top of head, close together; separate spinous and soft dorsal fins; dorsal and anal fins long-based; jaws large. About 12 species; marine, from shallow down to about 200 m; South America, Indo-Pacific to Japan.

Family Cheimarrichthyidae (torrent fish). Small, resembling a cottid or sculpin (family Cottidae); eyes on top of head and close together; 1 species; freshwater streams of New Zealand; young in brackish water.

Family Trachinidae (weever fishes). Eocene to present. Body elongated, compressed, deep at head end, tapering to narrow, small caudal fin; a separate short-based spinous dorsal fin with black membrane and poison glands along grooves in each spine; anal and soft dorsal fins long-based; a long spine on gill cover with poison glands along grooves of spine; eyes black in upper half, white in lower half; scales small, set in distinct oblique rows; 4 species, marine shallow water down to 90 m (300 ft); lie buried in sand with eyes and top of head showing; northern Europe, Mediterranean, Pacific coast of South America; venom very painful, even dangerous.

Family Trichonotidae (hairfins). Resemble Percophiidae and Mugiloididae, but body extremely elongated and dorsal fin unusually high; snout pointed; lips fringed, dive headfirst into sand. Several species; tropical and subtropical Indo-Pacific oceans.

Family Creediidae. Elongated little fishes resembling Percophiidae; 2 species known; coasts of Australia, Marshall and Mariana islands.

Family Limnichthyidae. Resemble Percophiidae; 5 species; New Zealand, Australia, and Society, Marshall, Mariana, and Hawaiian islands; sand burrowers.

Family Oxudercidae. A single species from coast of Asia at Macau; relationships still in doubt.

Superfamily Uranoscopoidea

Eocene to present. Three families; resembling toadfishes (Batrachoididae) and weever fishes (Trachinidae). Bodies rather elongated, with long dorsal and anal fins usually with few or no spinous rays; head broad, deep, flattish on top; eyes on top of head and somewhat erectile; mouth broad, oblique; lips fringed with toothlike dermal projections that interdigitate and strain water through sand when the fishes are buried to top of head and eyes; similar fringe of skin at upper end of gill slit; pectoral fins enlarged and fanlike; pelvic fins anterior to pectorals.

Family Leptoscopidae. Sand-burrowing fishes; no spines in dorsal and anal fins. Three species; marine; coasts of Australia and New Zealand; size up to 30 cm (12 in.).

Family Dactyloscopidae (sand stargazers). General features as described under superfamily; body elongated. Shape of pelvic fins is characteristic: each has 3 thickened segmented rays whose tips are free from the fin membrane, divergent and somewhat curled. Small fishes, mostly marine, up to about 10 cm (4 in.); burrow in sand with eyes and mouth protruding from surface of sandy bottom; 20 to 25 species in tropical America in Atlantic and Pacific oceans.

Family Uranoscopidae (stargazers). Head extremely broad and deep; posterior half of body tapering to a small truncate tail fin; eyes located on top of head, projecting above surface of head; large posteriorly pointing spine on shoulder girdle above pectoral fin (found in some species); dorsal fin moderately long and either with a short, nearly separate spinous portion or spines absent; anal fin moderately long and with few or no spines; large electric organs located behind eyes of species of genus *Astroscopus;* the only marine teleosts with electric organs; about 25 species live buried on sandy bottoms usually in shallow shore areas; a few species are found at greater depths, to about 600 m (2,000 ft), Atlantic, Pacific, and Indian oceans; size up to about 9 kg (20 lb).

Superfamily Chiasmodontoidea

Family Chiasmodontidae (swallowers). Slender fishes with extremely deeply cleft mouth; large, backward-pointing teeth; dorsal fin long with spinous and soft dorsals separate; pelvic fins thoracic. Capable of swallowing and holding in their greatly distensible bellies fishes larger than themselves. About 12 species in open oceanic waters down to 500 m (1,600 ft); size up to 15 cm (6 in.); relationships in doubt.

Superfamily Notothenioidea

Four families of codlike or sculpinlike percoids found mainly in the Antarctic, some in cold temperate Southern Hemisphere seas near Chile, Argentina, and New Zealand; superfamily includes about 75 percent of all Antarctic fishes.

Family Bovichthyidae. About 15 species in subantarctic and south temperate seas, off Chile, Argentina, southern New Zealand, and southern Australia; 1 species in rivers of South Australia and Tasmania.

Family Nototheniidae (Antarctic cods). Miocene to present; about 50 species, most in subantarctic waters; some species near Antarctic continent; a few in cold temperate zone, 1 species in rivers of southern South America. Mainly bottom dwellers of littoral zone, some deep-water species resemble true cods (Cottidae) and have a barbel on lower jaw; 2 species, large, up to 150 cm (5 ft).

Family Bathydraconidae (Antarctic dragonfishes). About 15 species; true Antarctic fishes, occurring on coasts of Antarctic continent; body greatly elongated; usually a spatulate, pikelike snout; no first dorsal fin; live on coasts of Antarctic continent to depths of 500–700 m, a few down to 2,500 m. Size up to about 50 cm, but usually much smaller.

Family Channichthyidae or Chaenichthyidae (white-blooded fishes or icefishes). Famous white-blooded fishes of the Antarctic; lack red blood cells and hemoglobin. Mostly large, up to 60 cm (2 ft) long, with scaleless body and 2 or 3 lateral lines each side; head large, snout long, spatulate, pikelike; teeth large; jaws nonprotractile. About 16 species, all in Antarctic except 1 in subantarctic; mainly bottom dwellers feeding on crustaceans and small fish; most at 100 to 200 m (330–660 ft), some to 700 m (2,300 ft).

Suborder Stromateoidei

Five percoidlike families with an unusual and characteristic feature, a toothed saccular outgrowth in the gullet directly behind the last gill arch. One family, the Amarsipidae, lacks the toothed saccular outgrowth in the gullet. Stromateoids may be related to fishes of the superfamily Kuhlioidea.

Families Stromateidae (butterfishes), *Centrolophidae* (medusafishes), *Nomeidae* (flotsamfishes), *Ariommidae, and Tetragonuridae* (squaretails). Eocene to present; slender to ovate, deep-bodied fishes; dorsal fin continuous or spinous portion set off from soft portion by deep notch; in the most generalized species, which resemble Kyphosidae, the soft dorsal is preceded by about 6 low, stoutish spines; other species resemble Carangidae. Eyes often with adipose (fatty) tissue around them; pelvic fins absent in some species, especially in ovate species; skeleton often weakly calcified; scaly sheath along bases of dorsal and anal fins. Young often found under jellyfishes or flotsam; adults live in deeper layers of ocean over continental shelves or are pelagic; a few species close to shore, in bays, even entering estuaries; 78 recognized species, many rarely seen; some feed largely on jellyfishes; others on crustaceans, tunicates, and small fishes; many are important commercially, especially in the Orient; family includes good-tasting species such as butterfishes and pompanos. Tropics to temperate zone; never near oceanic islands; size 10–120 cm.

Suborder Icosteoidei

Family Icosteidae (ragfish). A single species of rare deep-sea fish of North Pacific Ocean; body highly flexible in water, limp as a rag out of water; little is known of its anatomy. Resembles Stromateidae; presumably derived from a percoidlike ancestor; no spines in fins; pelvics and scales present in young, both absent in adult; body elongated, much compressed; up to 220 cm (7 ft).

Suborder Blennioidei

Fairly large suborder composed of 2 groups, the tropical blennies, some of which are percoidlike, and the northern blennies. The northern blennies are eel-like fishes, but the appearance of their faces and the aspects of their fins generally resemble those in the tropical blennies.

Superfamily Blennioidea (tropical blennies)

All soft, articulated fin rays in the dorsal, anal, and pelvic fins are unbranched; pelvic fins, which are seldom greatly reduced, consist of 1 spine and 4 or fewer rays and are located in front of the pectoral fins; there are the usual 2 nostrils on each side of the head. Exact 1-to-1 ratio between the vertebrae and the dorsal rays and posterior anal rays; only 1 lateral line. About 620 species.

Family Clinidae (clinids). Eocene to present. Percoidlike fishes, some moderately elongated, rather flat-sided, usually with somewhat pointed snouts and fleshy lips; dorsal and anal fins rather high and long-based, with fin membranes conspicuously supported by thin, riblike fin rays; caudal fin fanlike, not large; pelvic fins ahead of pectorals, slender; 1 spine and usually 2 or 3 rays; pelvics in some species appear usable in walking movements; cirri (bushy tentacles of skin) often present above eyes, on anterior nostrils, and just behind head on each side. Body scaled; scales cycloid; first 3 rays of dorsal fin often higher and more or less separate from rest of dorsal. A few species "four-eyed" (*i.e.,* with eye divided so as to see out of water). Most inhabit tide pools, kelp, rock crevices; some species down to 30 m (100 ft); size up to 30 cm (12 in.), but most are smaller; about 180 species.

Family Tripterygiidae (threefin blennies). Pliocene to present. Much like clinids but dorsal fin divided into 3 distinct parts, the first 2 of spines only; small bottom fishes of reef and rocks. About 100 species mostly in warm seas.

Family Blenniidae (combtooth blennies). Eocene to present. Resemble clinids in fins and body shape but differ in being scaleless, in having a steep forehead and only a single row of teeth in both jaws, the teeth being close-set, long, comblike. Sometimes a pair of large to enormous curved "canines" farther back in jaws; some species hop about out of water in intertidal zone; fins without pungent spines; eyes large and at top of forehead; pelvic fins with 2 rays. About 300 species, in tropical, subtropical, and warm temperate seas; small size.

Family Chaenopsidae (pike blennies). Pliocene to present. Body very elongated; jaws long; long gill area; dorsal and anal fins long, confluent with caudal fin; no scales or lateral line; 6 species in tropical and subtropical marine shore areas of Central America and Caribbean; small fishes living in worm tubes and burrows.

Family Pholidichthyidae (engineer fish). Very elongated eel-like fish; reclusive, living under excavations; move sand and gravel in mouths; known previously only by young stage; adult recently discovered by L. Dempster. One species; marine; in tropics, Indonesia and Philippines; size up to 30 cm (12 in.); poorly known; relationships in doubt.

Family Notograptidae. Eel-shaped; dorsal and anal fins long-based and high, both confluent with caudal fin; pelvics 1-rayed, filamentous, placed before pectorals; body scaled, mouth large. Two small species; marine; western Australia.

Superfamily Stichaeoidea (northern blennies)

Eel-like fishes; single pair of nostrils; dorsal and anal fins long-based and often joined to caudal fin; pelvic fins placed a little before pectoral fins, consisting of 1 spine and fewer than 4 rays; bottom-dwelling fishes usually of littoral zone, some supralittoral; a few in deep water (down to 450 m).

Family Stichaeidae (pricklebacks). Includes families Chirolophidae, Lumpenidae, Xiphiodontidae, Cebidichthyidae. Eel-like; body usually scaled; dorsal fin with spines only or some soft rays at rear of fin; pectorals reduced; pelvics present or absent; some species with 3 or 4 lateral lines but most species lacking lateral line. About 54 species, most in North Pacific, some in Atlantic; often among seaweeds when tide is out; also in deeper water (to 200 m); most feed on invertebrates, a few on seaweed; most species small, a few up to 60 cm.

Family Pholidae (gunnels). Elongated; strongly compressed laterally; scaled; pelvics reduced to 1 spine and 1 ray, very small or absent; spines only in dorsal fin. About 8 species, most in North Pacific; also in North Atlantic; size small; intertidal zone or shallow water; feed on small bottom invertebrates.

Family Anarhichadidae (wolffishes). Big head and long tapering body, laterally compressed; massive dentition on jaws, roof of mouth, and throat for crushing mollusks, sea urchins, crustaceans; length to 2.3 m (7.5 ft). Nine species, in northern oceans; littoral zone to 300 m (1,000 ft); good food fishes.

Family Ptilichthyidae (quillfish). Extremely elongated, body ending in a free, fleshy point; pelvic fins absent; dorsal and anal fins like vanes of a feather. One species, rare; North Pacific.

Family Congrogadidae. Includes Haliophidae. Resembles Stichaeidae but mouth and lips large; fins almost or entirely spineless; 3 lateral lines. About 8 species; rock-dwelling, in shallow coastal waters of Indo-Pacific from Africa to Japan.

Family Zaproridae (prowfish). A single species like a shorter, deeper bodied prickleback; pelvic fins absent; size up to 2.8 m (9 ft); deeper coastal waters to 350 m, California to Alaska.

Family Scytalinidae (graveldiver). Eel-like, with dorsal and anal fins soft-rayed and not beginning until middle of long straight body; body appears to flare out somewhat at these fins; pelvic fins lacking. One species; marine, California to Alaska; small, to 15 cm (6 in.); burrows quickly in sand or gravel in intertidal zone.

Suborder Acanthuroidei

Modified percoid-like fishes characterized by peculiarities of bones suspending the jaws, which thereby are extended far forward as small nibbling mouths at end of more or less lengthened snout.

Family Acanthuridae (surgeonfishes). Includes Zanclidae. Percoid-like fishes characterized by usually 1 or 2 lancet-like spines (like a surgeon's scalpel) alongside the caudal peduncle (in front of base caudal fin); body deep, compressed; dorsal and anal fins long-based; scales small; eyes high on side of head; mouth small, low, sometimes extended into a beak; teeth close-set and lobate. About 75 species; many herbivorous; inhabit shallow tropical shores, a few in deeper water.

Family Siganidae (rabbitfishes). Resembling surgeonfishes but uniquely characterized by pelvic fins each having 2 stout spines, located along the outer and inner edges, respectively; both spines are grooved for carrying venom; no lancet-like spine or buckler on caudal peduncle; mouth small and rabbitlike, used for nibbling algae. About 18 marine species of small to moderate size; tropical Indo-Pacific, around coral reefs and rocky areas.

Suborder Scombroidei

Streamlined, mackerel-like or marlin-like fishes the interrelationships of which are in doubt; upper jaw not protrusible; maxillary bones of upper jaw more or less firmly attached to non-protractile premaxillaries that lie ahead of them.

Family Gempylidae (snake mackerels). Eocene to present. Elongated, laterally compressed; mouth large, with large, cutting teeth; spinous part of dorsal fin longer than soft-rayed part, the latter often broken up into finlets posteriorly; pelvic fins usually not rudimentary. Sixteen species; tropical and temperate seas; down to 600 m (2,000 ft); length to 1 m (40 in.); some commercial value; good food fishes.

Family Trichiuridae (cutlass fishes). Paleocene to present. Elongated, bandlike body; large mouth with large fangs anteriorly. Scaleless; dorsal fin long, spinous anteriorly, anal fin long-based, spines short; caudal fin deeply forked on most species, absent in some (body ending in tapered point); pelvic fins reduced to scalelike spine and 1 small ray, or absent. Fifteen species in warm oceans; rather deep water offshore, to about 1,000 m (3,300 ft); piscivorous.

Family Scombridae (tunas and mackerels). Moderate to large, streamlined, swift-swimming, schooling fishes; body often thickly rounded, tapering to a narrow caudal peduncle bearing in some species 2 or 3 keels on its side; caudal fin widely forked or lunate (scimitar-shaped); distinguished from all fishes by series of separate finlets following spinous first dorsal and spinous anal fins; well-developed vascular system under skin of tunas is associated with sustained high-speed swimming and a body temperature a few degrees higher than that of the surrounding water; make extensive migrations to spawning and feeding grounds; fins fit into grooves or depressions on body; about 40 species, open waters of tropics and warm seas of world; feed on fish, crustaceans, squids, and other abundant animal life; length 30 cm to 3.3 m, weight to 800 kg. Many species subjected to ever-increasing fishing exploitation and will be threatened with extinction.

Family Xiphiidae (swordfish). Bones of upper jaw prolonged into a swordlike structure, flatter and sharp-edged, as compared with the round, shorter bill of marlins and billfishes; dorsal fin high and short-based, not extending beyond middle of body; caudal fin with high lobes; second dorsal fin small; pelvic fins absent; no teeth in jaws except in very young; no scales. Probably only 1 species, *Xiphias gladius*, worldwide in tropics and temperate seas; from surface to 400 m; size up to 3.6 m (141 in.) and 450 kg. (985 lb.). Sword is lashed about inside schools of fishes; injured fishes are eaten. A major big-game fish; excellent eating and commercially important.

Family Istiophoridae (billfishes, marlins, sailfishes, and spearfishes). Bill round and shorter compared with sword of swordfish; dorsal fin long, extending almost the length of back of body and reaching striking height in the sailfish, *Istiophorus gladius;* pelvic fins present as thin filaments; body scaled; 2 small keels on each side of caudal peduncle; fine teeth on jaws. About 7 species; worldwide in warm seas; all are large species, the black marlin (*Makaira indica*), the largest, probably attains 900 kg (2,000 lb); record for rod and reel is 710 kg (1,560 lb); greatest game fishes in the ocean, and excellent food fishes, of considerable economic importance.

Family Luvaridae (louvar). One species; rare; resembles a dolphin (family Coryphaenidae) in its very high forehead and eye placed low almost on level with mouth; mouth small, toothless; body deep, laterally compressed; a fleshy keel on each side of caudal peduncle; pelagic in tropics and subtropics; length to 1.5 m (5 ft).

Suborder Anabantoidei

Small percoid-like fishes characterized by an accessory air-breathing chamber of labyrinthic structure on each side of head, above regular gill chamber; an unusual patch of teeth on roof of mouth, teeth which attach to floor of cranium; usually build foam nests. Freshwater fishes of Southeast Asia, India, and Africa; about 70 species in 5 families, 4 of which are closely related; the fifth, Badidae, is more generalized.

Family Badidae. Resembling Anabantidae externally but lacking accessory air-breathing chamber. Have special patch of teeth on roof of mouth and characteristic courtship behaviour of anabantids; 1 species, freshwaters of India and Southeast Asia; small, 5–8 cm (2–3 in.) long.

Family Anabantidae. Pleistocene to present; about 20 species, freshwaters of tropical Africa, Southeast Asia and Philippines; includes the "climbing perch"; size small.

Family Belontiidae (fighting fishes, some gouramis, and others). About 50 species of small freshwater fishes from tropical Africa, India, Burma, Sri Lanka, Southeast Asia, Malaya; includes Siamese fighting fish (*Betta splendens*), various species of gouramis, which have 1 ray of each pelvic fin extended into an elongated filament; paradise fishes and other popular aquarium fishes. Most of these species were formerly placed in family Anabantidae.

Family Helostomatidae (kissing gourami). Freshwater, Southeast Asia; popular aquarium fish; one species.

Family Osphronemidae (gourami). Pelvic fins each with 1 ray drawn out into a long filament; 1 species; freshwater in Sumatra, Java, Borneo, grows to about 9 kg (20 lb).

Suborder Luciocephaloidei

Family Luciocephalidae (pikehead). Pike-like in body form and long head, an accessory suprabranchial air-breathing chamber (but not labyrinthic); bony plate in throat region; peculiar jaws. One species, freshwater; Malay Archipelago and Peninsula.

Suborder Mastacembeloidei

Family Mastacembelidae (spiny eels). Eel-like; head elongated with fleshy, mobile proboscis projecting beyond lower jaw; soft dorsal fin preceded by series of erectile spines. About 50 species, freshwaters from tropical Africa to Southeast Asia up to China; up to 90 cm (3 ft), but most much smaller; some attractively coloured aquarium fishes.

Family Chaudhuriidae. One species in freshwaters of Burma; similar to spiny eels but small and without fleshy proboscis.

Suborder Kurtoidei

Family Kurtidae (nurseryfishes). Peculiar, small, percoid-like; males carry eggs, stuck under an anteriorly pointing hornlike process on top of back of head. Two species; brackish water and lower parts of streams; Indo-Malaysia and New Guinea.

Suborder Echeneioidei

Family Echeneidae (remoras). Oligocene to present; differ from percoids mainly in having a sucking disk on top of head, modified from the spinous first dorsal fin. About 10 species; warm marine seas; ride attached to large fish or other marine animals or ships; carnivorous; some are cleaner fish for other fishes; length 30 to 90 cm (1 to 3 ft).

Suborder Ammodytoidei

Families Ammodytidae and Hypoptychidae (sand lances). Eocene to present. Long slender percoid-like fishes with long, pointed head and projecting lower jaw; dorsal fin long-based, with soft rays only; pelvic fins thoracic or (usually) absent. About 18 species; most seas of the world, especially colder waters; sand burrowers; large schools near shores form an important food source for many other fishes; 20–40 cm (8–16 in.).

Suborder Gobioidei

Almost all with pelvic fins located beneath pectorals and joined together to form a vacuum cup or suction disk; some with pelvic fins close together but not in form of a suction cup; a few lack pelvics; all lack one particular bone of cranial roof. More than 800 species, most of small size, living as bottom dwellers, worldwide, in saltwater, freshwater, or brackish water, especially in tropics.

Family Eleotridae (sleepers). Pelvic fins close together or in contact anteriorly, but not united into a sucking cup; short-based spinous first dorsal fin and longer based soft-rayed second dorsal; all fins usually with rather long fin rays; no lateral line. Numerous species; found along coasts and rivers in tropics and subtropics; most 10–30 cm (4–12 in.) long, one to 90 cm (35 in.) in length.

Family Gobiidae (gobies). Eocene to present. Pelvic fins united to form a suction cup; no lateral line; simple-looking but often colourful fishes with 2 dorsal fins; all fins rather conspicuously large compared with small, usually slender body; definite rows of naked lateral-line organs on head. More than 700 species; size from 1.2 cm (0.5 in., world's smallest vertebrates) to something under 30 cm (12 in.), with most less than 10 cm. Shallow coastal waters of tropics and temperate zones; usually resting on bottom or hidden; carnivorous.

Family Periophthalmidae (mud skippers). Elongated gobies with blunt head; eyes erectile, prominent, on top of head; pectoral fins with muscular base used to walk over mud and climb mangrove roots; about 6 species in estuaries and mud flats of Indian and Pacific oceans, to Japan.

Family Rhyacichthyidae (loach goby). Pelvic fins widely separated; head flattish, pointed; mouth ventral; 1 species living in torrential mountain streams of Indonesian Archipelago, up to Philippines; size up to 33 cm (13 in.).

Family Gobioididae. Elongated, eel-like gobies; dorsal fin very long, spinous part distinct from, but connected to, soft part; dorsal and anal fins confluent with caudal; eyes tiny to indistinct; mouth often obliquely upward; about 20 species, marine and brackish water of tropical America, Indian and Pacific oceans; about 8–30 cm (3–12 in.).

Family Trypauchenidae. Eel-like; scaled; dorsal and anal fins very long and confluent with caudal; pelvic fins united at base but not forming a cup or disk; eyes tiny. A blind cavity above gill cover opening to exterior by a pit. Burrowers in mud and gravel; about 12 species, along coasts and estuaries from Africa to Japan and Oceania.

Family Kraemeriidae. Rare little elongated fishes; pelvic fins separate; chin of lower jaw large, pointed, forming terminal end of head; eyes small, on top of head; 2 species, Indo-Pacific oceans.

Family Microdesmidae. Rare, small, eel-like; chin large, forming pointed end of snout; about 6 species; both coasts of tropical America, West Africa, tropical Pacific.

Suborder Callionymoidei

Family Callionymidae (dragonets). Eocene to present; resembling flatheads (Platycephalidae); body flattened, broad, muscular; snout very short; mouth small; gill openings restricted, located on top of head; eyes on top of head and close together; pelvics in front of pectorals; large spine pointing posteriorly at angle of preopercle on gill cover; bottom dwellers from tidepools down to 600 m (2,000 ft); burrow into sand; push along bottom using pelvic fins; about 40 species found in tropical and temperate zones.

Family Draconettidae. Look like callionymids but are separated on differences in head skeleton; no preopercular spine; a few species; North Atlantic and North Pacific in deep water.

Critical appraisal. Classification of perciform fishes will be receiving much more study in the future. Expected changes include removal of some groups from the order, addition of some from other orders, and considerable realignment of many groups within the Perciformes. Evidence suggests removal of Gobioidei to a preperciform position, possibly as a distinct order. The snake-heads (order Channiformes) should likely be returned to the order Perciformes, associated with the anabantoid and luciocephaloid fishes. Particularly difficult problems are the relationships and classifications of trachinoid, uranoscopoid, notothenioid, and stichaeoid fishes. Studies are presently being made on possible interrelationships between clusters within the 60 or more families of percoid fishes (suborder Percoidei). The placement and relationships of the atherinoid (sometimes called percesocine) fishes are still in dispute. The present classification follows that of Greenwood *et al.* in the removal of Atherinidae from perciform fishes and in the removal of the ophidioid and zoarcid fishes. Support for retaining the Atherinidae in order Perciformes and for other differences in the order Perciformes has been given by an American ichthyologist, W.A. Gosline, in a series of papers published between 1962 and 1971. Questions of relationships between carangid, rachycentrid, echeneid, and scombroid fishes are still unsettled. Many large and exciting problems thus remain. (W.C.F.)

Flatfishes—flounders, soles (Pleuronectiformes)

Members of the fish order Pleuronectiformes—which are commonly called flatfishes, flounders, or soles—are unique among fishes in being asymmetrical. Pleuronectiformes are strongly compressed, with both eyes on one side; other fishes and vertebrates in general are bilaterally symmetrical. The asymmetry is believed to have evolved from a generalized, symmetrical percoid (sea bass) body pattern in a fish that habitually rested on its side. Larval flatfishes have an eye on each side of the head, but during a period of rapid body change (metamorphosis) one eye migrates to the other side of the head, after which the larvae settle to the bottom. Osteological changes resulting from the eye migration are responsible for the asymmetry in the flatfish skull.

GENERAL FEATURES

Flatfishes of the family Pleuronectidae are commercially important in northern waters, and members of other families are taken in limited quantities. Some Bothidae and Soleidae (soles) are exploited in tropical and temperate waters, but no other flatfishes are utilized to the extent that Pleuronectidae are.

Flatfishes are primarily found in temperate and tropical seas, with some species extending northward into the Arctic. Sizes range from about 100 millimetres (four inches) to the large Atlantic halibut, which attains a length of more than two metres (nearly seven feet) and a weight of about 325 kilograms (716 pounds). Most species are marine, but some spend all or part of their life in freshwater. Flatfishes are found in depths up to 1,000 metres (3,300 feet), but most occur on the continental shelf in less than 200 metres of water.

By courtesy of (centre and bottom) U.S. Fish and Wildlife Service, from (top) S.F. Harmer and A.E. Shipley, *Cambridge Natural History* (1904), Macmillan & Co., London and Basingstoke

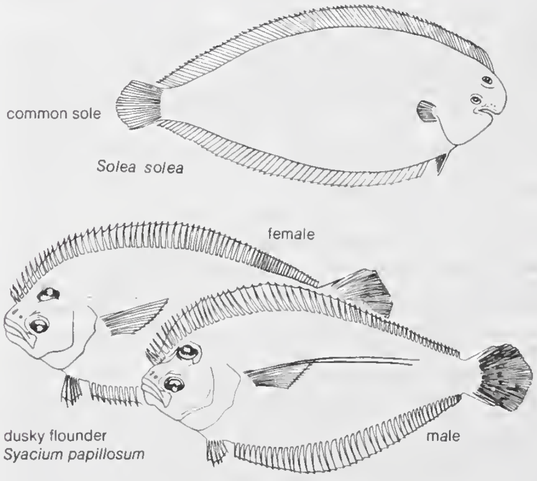


Figure 40: Body plan and sexual dimorphism in the flounder.

NATURAL HISTORY

Reproduction. Flatfishes generally spawn offshore, but some spawn in estuaries. Fecundity is high, females generally releasing at least several hundred thousand eggs (large female halibut have between 2,500,000 and 3,000,000 eggs). The eggs are small and float freely (pelagic) or sink to the botton (demersal), with or without oil globules. Newly hatched larvae are 1.5 to three millimetres long (approximately $^1/_{16}$ to $^1/_8$ inch). Active feeding begins shortly after hatching, and mortality of newly hatched larvae is extremely high. Larvae drift with currents (planktonic) until metamorphosis, or shortly afterward, and then settle to the bottom to assume their adult bottom-living (benthic) existence. Swimming is accomplished by undulating movements of body and fins, rather than by sculling with the caudal fin as in most other fish.

Feeding behaviour. Flatfishes lie on the bottom, generally covered by sand or mud, with only their eyes protruding. The eyes can be raised or lowered and moved independently. Flounders feed primarily on crustaceans, other bottom invertebrates, and small fish. When feeding they remain motionless until their prey ventures too close and then literally leap off the bottom in pursuit. Flatfishes in turn fall prey to a variety of large fish and cetaceans (whales, porpoises, etc.), but the primary predator of flatfishes is man.

FORM AND FUNCTION

Many species display sexual dimorphism, with the male having one or several of the following characteristics: elongated pectoral-fin rays, wider interorbital bones, spines on the head, tentacles on the eyes, more elaborate pigmentation. Flounders have a long dorsal fin extending from the head to caudal (tail) fin and an anal fin extending from vent (anus) to caudal fin in most species. Pectoral fins are present on all larval flatfishes but are lost or reduced in adults of the families Soleidae and Cynoglossidae. Caudal-fin rays and their supporting structures are variable. Scales are ctenoid (rough edged) or cycloid (smooth). Dentition is variable and corresponds to feeding habits of the species. Active predators have large, well-developed teeth in both jaws, whereas those living primarily in mud and feeding on bottom invertebrates have teeth only on the lower jaw of the blind side. Sexes are easily distinguished because the ovaries extend posteriorly from the body cavity beneath the skin and a thin layer of muscle immediately above the muscles of the anal-fin rays. Testes cannot be seen without dissection. The stomach and intestines curl within the body cavity to form a loop.

Migration of the eye during metamorphosis

The main feature of metamorphosis is the migration of the eye around or through the head. This is accomplished as a movement either over the middorsal ridge or through the head, in a depression between the supraorbital bars (over the eye) and ventral edge of the dorsal fin. The supraorbital bars extend forward from the cranium to the ethmoid region of the skull (the area in front of the eye), gradually shifting ventrally and coming to lie next to one another. As the eye migration begins, the dorsal edge of the supraorbital bar is reabsorbed to make room for the eye moving through the head. The supraorbital bars ossify and become the interorbital bone after the eye has completed its migration. The blind- (bottom-) side frontal bone shifts to the ocular (upper) side and forms a portion of the optic-capsule floor for the upper eye. Torsion (twisting) of the frontals, ethmoid, and mouthparts is the essential feature of the flatfish skull.

Normal pigmentation on adult flatfishes consists of a coloured ocular side and an unpigmented (white) blind side. The ocular side is variable in pigment pattern and intensity. Flatfishes can mimic their background by assuming a similar coloration. Partial or complete albinism is known in some species, but a more common colour variation is ambicoloration (coloration on both sides). Ambicoloration can be partial or complete and is often associated with incomplete migration of the eye (in which the migrating eye stops on middorsal ridge) and a hooked appearance, caused by the unattached origin of the dorsal fin. Reversal (eyes and pigmentation on the side that normally is unpigmented) is fairly frequent in some species but quite rare in others.

CLASSIFICATION

Flatfishes are divisible into three suborders and seven families.

ORDER PLEURONECTIFORMES
(Heterosomata of some authors)
Allied to Perciformes but asymmetrical, compressed, both eyes on one side of head; pelvic bones attached directly to cleithrum. Swim bladder absent in adults. Fossil records for this group of fish are limited, extending from Paleocene to Recent, about 65,000,000 years.

Suborder Psettodoidei
The least specialized (most primitive) flatfish. Spines present in dorsal, anal, and pelvic fins; dorsal fin not extending onto head; eyes on either right (dextral) or left (sinistral) side; maxillary (upper-jaw) bone with well-developed supplemental bone; vertebrae 24–25 (10 precaudal, 14–15 caudal).

Family Psettodidae (primitive flatfishes)
Same characters as given for the suborder. Two species, one from Indo-Pacific and one from Africa; attain a length of about 0.6 m (2 ft).

Suborder Pleuronectoidei
No spines in fins (one spine in pelvic fin of Citharidae); dorsal fin extending forward onto head; usually no supplemental bone on maxillary (may be present or absent in Citharidae); vertebrae 27–70 (generally numbering 34 or more); preopercular margin free; lower jaw prominent; nostrils asymmetrical (that on blind side being near edge of head).

Family Citharidae
Eyes either dextral or sinistral; anus on ocular side; gill membranes widely separated; dorsal- and anal-fin rays not shortened posteriorly. Five monotypic genera found in the Indo-Pacific and Mediterranean and off Africa and Japan; attain a length to about 30 cm (about 1 ft).

Family Scophthalmidae
Eyes sinistral; anus on blind side; gill membrane widely separated; dorsal- and anal-fin rays shortened posteriorly; pelvic-fin bases long (both extending forward onto the urohyal). Four genera and 10 species found in North Atlatic and Mediterranean Sea; attain lengths to about 1 m (3 ft 3 in.) and weights to about 23 kg (approximately 50 lb).

Family Bothidae (left-eyed flounders)
Eyes sinistral; anus generally far up on blind side; gill membranes connected; dorsal- and anal-fin rays shortened posteriorly; pelvic-fin bases on ocular side short, on blind side may be short or long, 6-fin rays in all but 1 species. Thirty genera with about 200 species, found primarily in the tropical and temperate seas of the world; generally small.

Family Pleuronectidae (right-eyed flounders)
Eyes dextral; anus on blind side, commonly on or near midline; gill membranes connected; dorsal- and anal-fin rays shortened posteriorly; pelvic-fin bases of ocular side short or long, on blind side short, 3–13 pelvic-fin rays. Forty-three genera with about 100 species, found primarily in northern and Arctic seas, but some occur in tropical and temperate seas.

Suborder Soleoidei
Same as Pleuronectoidei, but preopercular margin not free (covered by skin and scales); lower jaw not prominent; nostrils nearly symmetrical, that on blind side not near edge of head.

Family Soleidae (soles)
Eyes small, dextral; mouth curved downward; sensory or tactile papilla on head (primarily on blind side); caudal fin with numerous rays. Over 100 species found in tropic and temperate seas; many true soles are small and found along the shore, somes species inhabiting freshwater.

Family Cynoglossidae (tongue soles)
Eyes small, sinistral; head large; mouth curved downward, head lacking sensory or tactile papilla; dorsal and anal fins confluent with caudal fin; caudal fin with only about 12 fin rays; generally small, slender fishes, found in tropical seas.

(E.J.G.)

Box fishes, puffer fishes, ocean sunfishes, and allies (Tetraodontiformes)

The order Tetraodontiformes is a group of primarily tropical marine fishes that evolved from the Perciformes (the typical advanced spiny-rayed fishes) during the Eocene Period of the Cenozoic Era, about 50,000,000 years ago. The approximately 320 species of modern tetraodontiforms are notable for a high degree of diversity in anatomical structure and way of life. The great diversity evident among the 11 families of the order is also seen within some families, but not in others. Members of the deepwater, bottom-dwelling Triacanthodidae, the most primitive family, for example, range from relatively normal configurations to weirdly specialized forms with extremely long tubular snouts; the shallow-water members of the Triacanthidae, closely related and derived from the Triacanthodidae, are of rather uniform configuration. Likewise, the balistids are rather uniform in body plan; but monacanthids, which evidently evolved from them, include a series of species ranging from the normal to the exceedingly elongated and highly specialized.

GENERAL FEATURES

The tetraodontiforms make up about 5 percent of the tropical marine fishes of the world. Most species range in size from about eight to 60 centimetres (three to 24 inches) in length, but one ocean sunfish reaches more than three metres (11 feet). They are often strikingly patterned or gaudily coloured. With the exception of the relatively deepwater Triacanthodidae and Triodontidae, the members of this order are usually found in waters less than about 65 metres (200 feet) in depth and are especially prominent around coral or rocky reefs and on open sand and grass flats.

Poisonous flesh

Many species, especially of puffer fishes (Tetraodontidae), have poisonous flesh, at least during certain seasons of the year, but most of the highly poisonous substance (tetraodontoxin) responsible for the numerous annual fatalities in Indo-Pacific regions is contained in the viscera. The flesh of the poisonous species can be safely eaten only when the freshly caught specimen has been carefully cleaned and washed in the exacting manner of fugu (or puffer fish) chefs in Japan. But the majority of tetraodontiforms are palatable, and in numerous tropical regions the flesh of various triggerfishes and trunkfishes is highly esteemed. Other than as food in tropical coastal areas, man makes little direct use of tetraodontiforms, except for the dried bodies of the hard-cased boxfishes and the spine-studded, inflated puffers as curios. In fact, the order Tetraodontiformes contains so many strangely specialized species that the group has intrigued mankind from early times; a 1st-century Roman author, Pliny the Younger, for example, discussed puffer fishes and ocean sunfishes in his *Naturalis Historiae.* While most adult tetraodontiforms have thick, spiny skins or other defensive mechanisms that protect them from most predacious fishes, the young, relatively defenseless stages are eaten in great quantities by certain game fishes—dolphin, marlin and other billfishes, tunas, and various jacks.

From Spencer W Tinker, *Hawaiian Fishes* (1944)

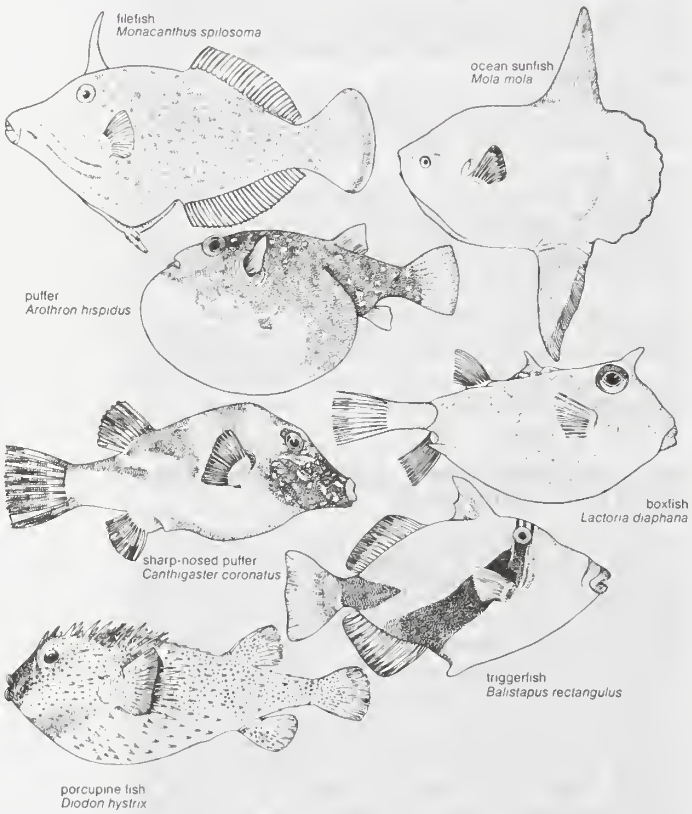


Figure 41: Body plans of representative tetraodontiform fishes.

NATURAL HISTORY

Feeding habits. As one would suspect from their usually well-developed and massive dentition, often with the teeth fused together in a parrotlike beak, most tetraodontiforms feed on hard-shelled crustaceans, mollusks, and echinoderms. But some, with massive, crushing jaws and teeth, such as the ocean sunfishes, often feed extensively on such soft-bodied invertebrates as jellyfishes (medusae). Some, such as boxfishes, blow a jet of water out of their mouths onto sand bottoms to expose burrowing invertebrates; others (such as some triggerfishes) specialize in eating spiny sea urchins or even clams and oysters. A few species, especially the long-snouted Triacanthodidae, have reduced or even rudimentary teeth, some apparently feeding on the scales of other bottom fishes. Other species probably feed on soft-bodied invertebrates, probing with the snout into holes in the bottom or into recesses in outcroppings to obtain food unavailable to less specialized fishes. Although many species have specialized feeding habits, the order as a whole can be considered as comprising opportunistic predators on invertebrates.

Locomotion. Most tetraodontiforms swim by the rather unusual method of rapid undulations or complex scullings of the soft dorsal and anal fins (in the midline of the back and underside, respectively); the powerful caudal fin (except in the Molidae) is reserved for rapid bursts of speed. The paired pectoral fins (just behind the gills) are in an almost constant state of rapid vibration, which gives a delicacy of control to their movements that is unusual even among fishes.

Activity cycle. Those tetraodontiforms for which data are available are diurnal, feeding or otherwise active during daylight but quiescent at night, often retiring to holes or crevices in coral or rocky reefs to sleep. When disturbed during the day, as by a potential predator, some species take rapid flight; others dive into reef crevices. Other species avoid the attention of predators by remarkable colorations or patternings that permit them to blend into the environment, which may be anything from a coral reef to a bed of bottom sea grass. One relatively defenseless species (a filefish), moreover, is an excellent mimic in body form and bright coloration of a spiny-skinned, inflatable, and perhaps poisonous puffer.

FORM AND FUNCTION

The Tetraodontiformes are distinguished externally by a small gill opening restricted to a relatively short slit on the side of the head and a small mouth, usually equipped with massive teeth. The scales of the body are typically highly modified into overlapping (in triacanthoids and balistoids) and even sutured (in ostraciontoids) plates or into sharp, projecting spines (in tetraodontoids and diodontids); in some cases, the skin itself may be thickened and hardened by deep layers of connective tissues (molids). There are no anal fin spines, and the dorsal fin spines are either absent or present only in reduced number (never more than six). The pelvic fin, which in the perciforms has one spine and five soft rays, in tetraodontiforms is either absent or reduced to no more than one spine and two small soft rays. The skeleton of tetraodontiforms is notable for a reduced number of bones, a number of the separate bony elements of the ancestral perciforms having been lost through the processes of reduction, consolidation, fusion, or failure to develop. The hallmark of the evolution and diversification of the tetraodontiforms, in fact, has been the reductive tendencies in some parts—number of skeletal elements, number of fin spines, size of mouth and gill opening, and number of teeth—with the simultaneous elaborative tendencies in other systems—scale and skin development, inflation apparatus, size of teeth and fusion with jawbones, and poisonous flesh.

Fin and skeletal features

CLASSIFICATION

Annotated classification. The Tetraodontiformes are classified as follows, with only the most obvious external differences that distinguish the groups mentioned.

ORDER TETRAODONTIFORMES (PLECTOGNATHI)

Small mouth and gill openings; reduced dorsal and pelvic fin spines; no anal fin spines; skin usually tough or spiny.

Suborder Balistoidei (Sclerodermi)

Teeth separate, discrete, individual units.

Superfamily Triacanthoidea

Six dorsal spines and a large pelvic spine.

Family Triacanthodidae (spikefishes). The most primitive members of the order; deepwater species with a truncated or rounded tail; deep caudal peduncle (the region between the end of the anal fin and the front of the tail); nonstreamlined body; soft dorsal and anal fins of about same length along their bases; Indo-Pacific and Caribbean.

Family Triacanthidae (triple spines). Shallow-water derivatives of the spikefishes; deeply forked caudal fin; slender caudal peduncle; body relatively streamlined for rapid swimming; soft dorsal fin base much longer than anal fin base; Indo-Pacific, sometimes found in estuaries.

Superfamily Balistoidea

Two or 3 dorsal spines, the 2nd spine serving to lock the 1st in an erected position; pelvic spine rudimentary or absent.

Family Balistidae (triggerfishes). Three dorsal spines; 8 outer teeth in each jaw; worldwide.

Family Monacanthidae (filefishes). Two dorsal spines; 6 or fewer outer teeth in each jaw; worldwide.

Superfamily Ostraciontoidea

No dorsal spines, body encased in a turtle-like cuirass (carapace) of sutured, platelike scales.

Family Aracanidae (keeled boxfishes). Carapace open behind the dorsal and anal fins and bearing a ventral keel; usually in deeper water than the ostraciontids; Indo-Pacific.

Family Ostraciontidae (boxfishes, trunkfishes, cowfishes). Carapace closed behind anal and usually behind dorsal fin, no ventral keel; worldwide.

Suborder Tetradontoidei (Gymnodontes)

Teeth more or less fused to the jawbones, forming a parrotlike beak.

Superfamily Triodontoidea

Three tooth plates, 2 in upper and 1 in lower jaw.

Family Triodontidae (pursefish). Most primitive member of the suborder, the only species to retain even the pelvic bone of the pelvic fin apparatus (completely lost by all other members of suborder); body somewhat elongate; 1 species; deep water; tropical Indo-Pacific.

Superfamily Tetraodontoidea

Four tooth plates, 2 in each jaw; the skin bearing small erectile spines.

Family Tetraodontidae (puffer fishes). A large number of species, differing from the sharp-nosed puffers mainly in osteological structure, but always having a prominent nasal apparatus; worldwide.

Family Canthigasteridae (sharp-nosed puffer fishes). Single, inconspicuous nostril on each side of head; snout more laterally compressed than in the tetraodontids; worldwide.

Superfamily Diodontoidea

Two tooth plates, 1 in each jaw; the skin bearing huge spines; caudal fin normal.

Family Diodontidae (porcupine fishes and burrfishes). Characteristics of superfamily. Spines erectile (porcupine fishes) or fixed (burrfishes); worldwide.

Superfamily Moloidea

Two tooth plates, 1 in each jaw. Skin relatively smooth but often exceptionally thick; caudal fin highly modified or absent; swim bladder absent.

Family Molidae (ocean sunfishes). Three species, 2 of which reach enormous size, 1 up to 3.3 metres (11 feet) in length and 1,900 kilograms (4,000 pounds) in weight; tropical and subtropical oceans worldwide.

Critical appraisal. The classification of the order is still in a state of flux, and some authorities are willing to recognize as families only those that are listed above as superfamilies. More important, the ostaciontoids are often recognized as a third suborder of Tetraodontiformes rather than, as above, a superfamily of the suborder Balistoidei. Only recently has the ordinal term Tetraodontiformes come into use; the group still is often called the Plectognathi and its two suborders the Sclerodermi and Gymnodontes. (J.C.T.)

BIBLIOGRAPHY

Fishes. *General works:* E.S. HERALD, *Living Fishes of the World* (1961), a clearly written and extensively illustrated introduction to fishes; G.U. LINDBERG, *Fishes of the World* (1971; Eng. trans. from the Russian, 1974), a comprehensive work with extensive bibliography; J.S. NELSON, *Fishes of the World* (1976), a treatment of all families that includes maps showing distribution; P.P. GRASSÉ (ed.), *Traité de zoologie*, vol. 13, *Agnathas et Poissons*, 3 parts (1958), a classic and authoritative review in French of the classification, anatomy, and biology of fishes. J.R. NORMAN, *A History of Fishes*, 3rd ed. by P.H. GREENWOOD (1975); N.B. MARSHALL, *The Life of Fishes* (1965); K.F. LAGLER *et al.*, *Ichthyology*, 2nd ed. (1977), college-level introductory texts of general ichthyology; J.E. WEBB, J.A. WALLWORK, and J.H. ELGOOD, *Guide to Living Fishes* (1981); and TIM M. BERRA, *An Atlas of Distribution of Freshwater Fish Families of the World* (1981).

Regional works: A.H. LEIM and W.B. SCOTT, "Fishes of the Atlantic Coast of Canada," *Bull. Fish. Res. Bd. Can.*, no. 155 (1966), a good general account, completely illustrated; H.B. BIGELOW *et al.*, *Fishes of the Western North Atlantic*, 5 vol. (1948–66), a comprehensive treatment of the biology of western North Atlantic fishes; J.E. BÖHLKE and C.C.G. CHAPLIN, *Fishes of the Bahamas and Adjacent Tropical Waters* (1968); W.B. SCOTT, *Freshwater Fishes of Canada* (1973); W.A. CLEMENS and G.V. WILBY, "Fishes of the Pacific Coast of Canada," *Bull. Fish. Res. Bd. Can.*, no. 68, 2nd ed. (1961); J. and G. LYTHGOE, *Fishes of the Sea: The Coastal Waters of the British Isles, Northern Europe and the Mediterranean* (1971); W.A. GOSLINE and V.E. BROCK, *Handbook of Hawaiian Fishes* (1960), an excellent handbook of fishes from the central Pacific Ocean; T.C. MARSHALL, *Fishes of the Great Barrier Reef and Coastal Waters of Queensland* (1964); T. KAMOHARA, *Fishes of Japan in Color* (1967); J.T. NICHOLAS, *The Fresh-Water Fishes of China* (1943), somewhat old, but a complete coverage on fishes of eastern Asia; I.S.R. MUNRO, *The Marine and Freshwater Fishes of Ceylon* (1955), an inclusive illustrated account of fishes from the Indian Ocean area; J.L.B. SMITH, *The Sea Fishes of Southern Africa*, 5th ed. (1965); R.H LOWE-MCCONNELL, *Fish Communities in Tropical Freshwaters: Their Distribution, Ecology and Evolution* (1975), a broad review of fishes of Africa, South America, and Asia. Somewhat more local, but applicable to much of North America, are the following: M.B. TRAUTMAN, *The Fishes of Ohio with Illustrated Keys*, rev. ed. (1980); F.B. CROSS, *Handbook of Fishes of Kansas* (1967), and, with J.T. COLLINS, a companion vol., *Fishes in Kansas* (1975); H.D. HOESE and R.H. MOORE, *Fishes of the Gulf of Mexico: Texas, Louisiana and Adjacent Waters* (1977); W.F. SMITH-VANIZ, *Freshwater Fishes of Alabama* (1968); C.L. HUBBS and C. LAGLER, *Fishes of the Great Lakes Region*, rev. ed. (1958); JAMES F. MORROW, *The Freshwater Fishes of Alaska* (1980); and ROBERT J. NAIMAN and DAVID L. SOLTZ (eds.), *Fishes in North American Deserts* (1981).

Natural history: C.M. BREDER and D.E. ROSEN, *Modes of Reproduction in Fishes* (1966), a summary of reproductive behaviour of fishes; N.B. MARSHALL, *Aspects of Deep Sea Biology* (1954), a college-level introduction to deep-sea biology; B.W. HALSTEAD, *Poisonous and Venomous Marine Animals of the World*, vol. 2 and 3, rev. ed. (1978), an extensive treatment of poisonous and venomous marine fishes, beautifully illustrated in colour; H.S. DAVIS, *Culture and Diseases of Game Fishes* (1953), a general aid to the culture of North American game fishes. See also MICHAEL GOULDING, *The Fishes and the Forest. Explorations in Amazonian Natural History* (1981).

Form and function: R.M. ALEXANDER, *Functional Design in Fishes*, 3rd ed. (1974), a short college-level book on functional anatomy of fishes; M.E. BROWN (ed.), *The Physiology of Fishes*, 2 vol. (1957); and W.S. HOAR and D.J. RANDALL (eds.), *Fish Physiology*, 6 vol. (1969–71), advanced general texts; NATO, *Environmental Physiology of Fishes* (1980).

Paleontology and classification: L.S. BERG, *System der rezenten und fossilen Fischartigen und Fische* (1958), a German translation from the second Russian edition, a revised edition of Berg's 1940 classification of contemporary and fossil fishes; A.S. ROMER, *Vertebrate Paleontology*, 3rd ed. (1966), a college-level text with a good review of fish evolution; W.A. GOSLINE, *Functional Morphology and Classification of Teleostean Fishes* (1971), a study of evolutionary relationships among orders of teleost fish.

Bibliographic reference: B. DEAN *et al.* (eds.), *A Bibliography of Fishes*, 3 vol. (1916–23, reprinted 1972), an almost complete bibliography of works on contemporary and fossil fishes up to about 1923.

Agnatha. V.C. APPLEGATE, *Natural History of the Sea Lamprey, Petromyzon marinus, in Michigan*, Spec. Scient. Rep. U.S. Fish Wildl. Serv. 55 (1950), an account of lamprey life history and the pest problem; A. BRODAL and R. FANGE (eds.), *The Biology of Myxine* (1963), an authoritative work by 24 authors; M. EONTAINE, "Formes actuelles de cyclostomes," in P.P. GRASSÉ (ed.), *Traité de zoologie,* vol. 13, pp. 1–172 (1958), the best general account of cyclostomes (well illustrated); M.W. HARDISTY and I.C. POTTER (eds.), *The Biology of Lampreys* (1972), a comprehensive reference; A.J. MARSHALL, "Agnatha," in T.J. PARKER and W.A. HASWELL, *Textbook of Zoology,* 7th ed., vol. 1 (1962), a good account of comparative anatomy, written for university students; E. STENSIO, "Les Cyclostomes fossiles ou ostracodermes," in J. PIVETEAU (ed.), *Traité de paléontologie,* vol. 4, pp. 96–382 (1964); R. STRAHAN, "The Behaviour of Myxinoids," *Acta Zool. Stockh.,* 44:73–102 (1963), the most recent accounts of the subject.

Chondrichthyes (Selachii). J.S. BABEL, "Reproduction, Life History, and Ecology of the Round Stingray, *Urolophus halleri* Cooper," *Fish Bull. Calif. 127* (1967), an excellent natural history study of a ray; H.B. BIGELOW and W.C. SCHROEDER, "Sharks," *Fishes of the Western North Atlantic, Mem. Sears Fdn. Mar. Res.,* no. 1, pt. 1, pp. 59–546 (1948); "Sawfishes, Guitarfishes, Skates and Rays," and "Chimaeroids," *ibid.,* no. 1, pt. 2 (1953), the two most recent cosmopolitan syntheses in spite of emphasis on the western North Atlantic; "A Study of the Sharks of the Suborder Squaloidea," *Bull. Mus. Comp. Zool. Harv.,* vol. 117, no. 1 (1957), a revision of the group; P. BUDKER, *La Vie des requins* (1947; rev. Eng. trans., *The Life of Sharks,* 1971), a nontechnical, authoritative work; J.E. DANIEL, *The Elasmobranch Fishes,* 3rd rev. ed. (1934), a classic treatise on the anatomy of sharks and rays; D.H. DAVIES, *About Sharks and Shark Attack* (1964), an authoritative, nontechnical work about sharks, principally South African species; P.W. GILBERT (ed.), *Sharks and Survival* (1963), a collection of 22 papers about sharks, with emphasis on dangerous species, including a list of documented attacks throughout the world; P.W. GILBERT, R.E. MATHEWSON, and D.P. RALL (eds.), *Sharks, Skates and Rays* (1967), a collection of technical papers covering a wide range of subjects; E.W. GUDGER (ed.), *Archaic Fishes,* Bashford Dean Memorial Volume (1937), contains articles on the anatomy, reproduction, and development, of the frilled shark (*Chlamydoselachus anguineus*), and on natural history and development of Heterodontid sharks; H.W. MCCORMICK and T. ALLEN, with W.E. YOUNG, *Shadows in the Sea: The Sharks, Skates, and Rays* (1963), an authoritative, well-illustrated, nontechnical work; L.H. MATTHEWS and H.W. PARKER, "Notes on the Anatomy and Biology of the Basking Shark (*Cetorhinus maximus*)," *Proc. Zool. Soc. Lond.,* 120:535–576 (1950); P.M. ROEDEL and W.E. RIPLEY, "California Sharks and Rays," *Fish Bull. Calif. 75* (1950), a handbook for identification of California species (of general use on the Pacific coast), illustrated with excellent photographs; A.S. ROMER, *Vertebrate Paleontology,* 3rd ed. (1966), a comprehensive review, with a classification of archaic and recent fishes; H.W. SMITH, *From Fish to Philosopher* (1961), a chapter on the elasmobranchs discusses their evolution and physiological adaptations to marine environment; D.W. STRASBURG, "Distribution, Abundance, and Habits of Pelagic Sharks in the Central Pacific Ocean," *Fish Bull. U.S. Dep. Interior,* 58:335–416 (1958), contains field observations and quantitative data on 12 species; G.P. WHITLEY, *The Fishes of Australia, Part I, The Sharks, Rays, Devil-Fish, and Other Primitive Fishes of Australia and New Zealand* (1940), a nontechnical book about Australian species; CARL GAUS and T.S. PARSONS, *A Photographic Atlas of Shark Anatomy: The Gross Morphology of 'Squalus Acanthias'* (1981).

Crossopterygii. There is virtually no popular literature dealing with this group of fishes. The following are technical in nature: E. JARVIK, "On the Structure of the Snout of Crossopterygians and Lower Gnathostomes in General," *Zool. Bidr. Upps.,* 21:235–675 (1942); J.P. LEHMAN, "Crossopterygii," in *Traité de paléontologie IV,* 3:301–412 (1966); J. MILLOT, and J. ANTHONY, "*Latimeria chalumnae,* dernier des Crossoptérygiens," in *Traité de zoologie,* vol. 13, pp. 2553–2597 (1958); *Anatomie de* Latimeria chalumnae, 2 vol. (1958–66); J.L.B. SMITH, "A Living Coelacanthid Fish from South Africa," *Trans. R. Soc. S. Afr.,* 28:1–106 (1940); K.S. THOMSON, "The Comparative Anatomy of the Snout in Rhipidistian Fishes," *Bull. Mus. Comp. Zool. Harv.,* 131:313–357 (1964).

Dipnoi. H. SWAN, D. JENKINS, and K. KNOX, "Metabolic Torpor in *Protopterus aethiopicus:* An Anti-Metabolic Agent from the Brain," *Am. Nat.,* 103:247–258 (1969), an article on dry sleep in lungfishes; M. BLANC, E. D'AUBENTON, and Y. PLESSIS, "Mission M. Blanc-F. d'Aubenton (1954) IV. Ètude de l'enkystement de *Protopterus annectens* (Owen 1839)," *Bull. Inst. Fr. Afr. Noire,* Series A, 18:843–854 (1956), a study of the encystment of *Protopterus annectens;* P. BRIEN, M. POLL, and J. BOUILLON, "Ethologie de la reproduction du *Protopterus dolloi* (Boulenger)," *15th Int. Congr. Zool.,* sect. 1 (1959); J.S. BUDGETT, "On the Breeding-Habits of Some West-African Fishes, with an Account of the External Features in the Development of *Protopterus annectens,* and a Description of the Larva of *Polypterus lapradei,*" *Trans. Zool. Soc. Lond.,* 16:115–136 (1901); K. CURRY-LINDAHL, "On the Ecology, Feeding Behaviour and Territoriality of the African Lungfish, *Protopterus aethiopicus,* Heckel," *Ark. Zool.,* Series 2, 9:479–497 (1956); A.G. JOHNELS and G.S.O. SVENSSON, "On the Biology of *Protopterus annectens* (Owen)," *ibid.,* 7:131–164 (1955), a detailed study of the habits of lungfishes in the flooding zones on both sides of the Gambia River; K.H. LUELING, "Einige Notizen uber afrikanische Lungenfische," *Dt. Aquar.-Terrar-Z.,* 12:12–14, 44–46 (1959), on the habits and distribution of the African lungfishes, together with a distribution map according to Poll; "Untersuchungen an Lungenfischen, insbesondere an afrikanischen Protopteriden," *Bonn. Zool. Beitr.,* 12:87–112 (1961), a detailed examination of the experimental encysting of the West African lungfish *Protopterus dolloi* in captivity; "Fische mit Lungen," *Neptun,* 6:80–83 (1966), a study of the morphology, anatomy, and the method of breating of lunglike structures in the dipnoi; M. POLL, "Zoogéographie des protoptères et des polyptères," *Bull. Soc. Zool. Fr.,* 79:282–289 (1955), a discussion of the distribution of the four species of African lungfishes, with distribution map; H.W. SMITH, "Metabolism of the Lung-Fish, *Protopterus aethiopicus,*" *J. Biol. Chem.,* 88:97–130 (1930), the first modern physiological study of the encysting of Ethiopian lungfishes in captivity; "Observations on the African Lung-Fish, *Protopterus aethiopicus,* and on Evolution from Water to Land Environments," *Ecology,* 12:164–181 (1931); E.K. SUVOROV, *Allgemeine Fischkunde* (1959; German trans. from the 2nd Russian ed. of 1948), includes a chapter on the breathing organs of Dipnoi.

Chondrostei. S.M. ANDREWS *et al.,* "Pisces," in W.B. HARLAND *et al.* (eds.), *The Fossil Record: A Symposium with Documentation,* ch. 26 (1967), a recent classification of fish, with first and last occurrences for each family; E.S. GOODRICH, "*Vertebrata craniata,*" fasc. 1, "Cyclostomes and Fishes," in E.R. LANKESTER (ed.), *A Treatise on Zoology* (1909), a classic work on the anatomy of fish that is still useful; D. HEYLER, *Vertébrés de l'autunien de France* (1969), a detailed account of some palaeonisciform fish from a classic locality in France; J.P. LEHMAN, "Etude complementaire des poissons de l'Eotrias de Madagascar," *K. Svenska Vetensk-Akad. Handl.,* 2:1–201 (1952), a detailed study of early Triassic fish from Madagascar; "Superordre des Chondrostéens (Chondrostei): Formes fossiles," in P.P. GRASSÉ (ed.), *Traité de zoologie,* vol. 13 (1958), a comprehensive reference work on the organization and classification of the chondrosteans, and "Actinopterygii," in J. PIVETEAU (ed.), *Traité de paléontologie,* vol. 4 (1966); D.V. OBRUCHEV (ed.), *Fundamentals of Paleontology,* vol. 11, *Agnatha, Pisces* (1967); and A.S. ROMER, *Vertebrate Paleontology,* 3rd ed. (1966), two comprehensive reference works.

Holostei. S.M. ANDREWS *et al.,* "Pisces," in W.B. HARLAND *et al.* (eds.), *The Fossil Record: A Symposium with Documentation,* ch. 26 (1967), a recent classification of fish, with first and last occurrences for each family; B.G. GARDINER, "A Revision of Certain Actinopterygian and Coelachanth Fishes, Chiefly from the Lower Lias," *Bull. Br. Mus. Nat. Hist.* (*Geol.*), 4:239–384 (1960), important revised descriptions of early Jurassic fish from Great Britain; E.S. GOODRICH, "*Vertebrata craniata,*" fasc. 1, "Cyclostomes and Fishes," in E.R. LANKESTER (ed.), *A Treatise on Zoology* (1909), a classic work on the anatomy of fish that is still useful; J.P. LEHMAN, "Actinopterygii," in J. PIVETEAU (ed.), *Traité de paléontologie,* vol. 4 (1966), a recent summary of important characteristics of the higher bony fishes, along with their geologic and geographic distributions; D.V. OBRUCHEV (ed.), *Fundamentals of Paleontology,* vol. 11, *Agnatha, Pisces* (1967), a summary treatment of all fishes, living and fossil; D.H. RAYNER, "The Structure and Evolution of the Holostean Fishes," *Biol. Rev.,* 16:218–237 (1941), an attempt to relate the various families of holostean fishes mainly on the basis of braincase design; A.C. WEED, *The Alligator Gar* (1923).

Teleostei. For references to the literature on teleostei, see the following bibliographies for the major teleost groups.

Elopiformes. S.E. HILDEBRAND, "Family Elopidae" and "Family Albulidae," in H.B. BIGELOW *et al., Fishes of the Western North Atlantic,* vol. 3, pp. 111–147 (1963), full discussions of the structure, biology, and classification of ladyfishes, tarpons, and bonefishes; R.A. WADE, "The Biology of the Tarpon, *Megalops atlanticus,* and the Ox-eye, *Megalops cyprinoides,* with Emphasis on Larval Development," *Bull. Mar. Sci. Gulf Caribb.,* 12:545–622 (1962), a detailed account of the biology of the Pacific tarpon, or oxeye, and the Atlantic tarpon; P.H. GREENWOOD, "Skull and Swimbladder Connections in Fishes of the Family Megalopidae," *Bull. Br. Mus. (Nat. Hist.), Zool.,* 19:119–135 (1970), details of the specialized respiratory system of the tarpons; P.L. EOREY, "A Revision of the Elopiform Fishes, Fossil and Recent," *Bull. Br. Mus. (Nat. Hist.), Geol.,* suppl. 10 (1973), a full account of the structure and evolution of elopiforms.

Anguilliformes. L. BERTIN, *Les anguilles,* 2nd ed. (1951; Eng. trans., *Eels: A Biological Study,* 1956), a comprehensive, readable account of the biology of freshwater eels; J. SCHMIDT, "The

Breeding Places of the Eel," *Rep. Smithson. Instn. (1924)*, pp. 279–316 (1925), a summary by the original researcher of the classical biological study of eels; A.E. BRUUN, "The Breeding of the North Atlantic Freshwater-Eels," *Adv. Mar. Biol.*, 1:137–170 (1963), an important review of current controversy regarding the eel life-cycle; L. BERTIN and C. ARAMBOURG, "Anguilliformes," in P.P. GRASSÉ (ed.), *Traité de Zoologie*, 13:2314–2327 (1958), a classical text in French with a good coverage of eels; P.H.J. CASTLE, "The World of Eels," *Tuatara*, 16:85–97 (1968), a general article of college level difficulty, and "An Index and Bibliography of Eel Larvae," *Spec. Publs. Inst. Ichthyol. Rhodes Univ.*, 7:1–125 (1969), a comprehensive bibliographic index to leptocephali; F.W. TESCH, *The Eel: Biology and Management of Anguillid Eels* (1977).

Clupeiformes. P.H. GREENWOOD *et al.*, "Phyletic Studies of Teleostean Fishes with a Provisional Classification of Living Forms," *Bull. Am. Mus. Nat. Hist.*, 131:339–455 (1966), contains the latest classification, which is used also in this article. Some different classifications and arrangements are in: L.S. BERG, "Classification of Fishes, Both Recent and Fossil," *Trav. Inst. Zool. Acad. Sci. U.S.S.R.*, vol. 5, no. 2 (1940; reprinted in book-form, 1947), Russian and English texts; LEON BERTIN and CAMILLE ARAMBOURG, "Super-ordre des téléostéens (Teleostei)," in P.P. GRASSÉ (ed.), *Traité de zoologie*, vol. 13, fasc. 3, pp. 2204–2500 (1958); P.J.P. WHITEHEAD, "A Contribution to the Classification of Clupeoid Fishes," *Ann. Mag. Nat. Hist.*, Series 13, 5:737–750 (1962); H.S. CLAUSEN, "Denticipitidae, a New Family of Primitive Isospondylous Teleosts from West African Freshwater," *Vidensk. Meddr. Dansk Naturh. Foren.*, 121:141–151 (1959), the first scientific description of an unknown and evolutionary important Clupeiformes family; M.B. SCHAEEER, "Men, Birds and Anchovies in the Peru Current: Dynamic Interactions," *Trans. Am. Fish. Soc.*, 99:461–467 (1970), a modern evaluation of the most abundant species resources; A.J. MANSUETI and J.D. HARDY, *Development of Fishes of the Chesapeake Bay Region: An Atlas of Egg, Larval, and Juvenile Stages* (1967); E.K. BALON, "First Catches of Lake Tanganyika Clupeids (kapenta—*Limothrissa miodon*) in Lake Kariba," *Bull. Fish. Res. Zambia*, 5:175–186 (1971), an account of successful artificial introduction into a new area; and A.N. SVETOVIDOV, "Fauna of the U.S.S.R.," *Fishes*, vol. 2, no. 1, *Clupeidae* (1963; originally published in Russian, 1952), a good source of biological and morphological data with extensive bibliography.

Osteoglossomorpha. P.H. GREENWOOD, "On the Genus *Lycoptera* and Its Relationship with the Family Hiodontidae (Pisces, Osteoglossomorpha)," *Bull. Br. Mus. Nat. Hist (Zool.)*, 19:259–285 (1970); and *et al.*, "Phyletic Studies of Teleostean Fishes, with a Provisional Classification of Living Forms," *Bull. Am. Mus. Nat. Hist.*, 131:339–455 (1966), includes the most recent classification of the Osteoglossomorpha and a discussion of the reasons for that arrangement; E.S. HERALD, *Living Fishes of the World* (1961), a popular, well-illustrated account of the various families, and some species, particularly their biology; K.H. LULING, "*Arapaima*, Giant Fish of Amazonas," *Animals*, 11:222–225 (1968), a popular account; G.J. NELSON, "Infraorbital Bones and Their Bearing on the Phylogeny and Geography of Osteoglossomorph Fishes," *Am. Mus. Novit.*, no. 2394 (1969); "Gill Arches of Teleostean Fishes of the Division Osteoglossomorpha," *J. Linn. Soc. (Zool.)*, 47:261–277 (1968); J.R. NORMAN, *A History of Fishes*, 2nd ed. rev. by P.H. GREENWOOD (1963), includes a general account of osteoglossomorph biology, distribution, and anatomy (high school and college level); G. STERBA, *Susswasserfische aus aller Welt* (1959; Eng. trans., *Freshwater Fishes of the World*, 1962), particularly for the aquarist.

Salmoniformes. J.W. JONES, *The Salmon* (1959); W.E. EROST and M.E. BROWN, *The Trout* (1967), two works with general information on Salmoniformes; J.E. EITCH and R.J. LAVENBERG, *Deep-Water Teleostean Fishes of California* (1968), a book designed for the interested layman, covering many deep-sea Salmoniformes; P.H. GREENWOOD *et al.*, "Phyletic Studies of Teleostean Fishes with a Provisional Classification of Living Forms," *Bull. Am. Mus. Nat. Hist.*, 131: 339–455 (1966), created the order Salmoniformes; S.H. WEITZMAN, "The Origin of the Stomiatoid Fishes with Comments on the Classification of Salmoniform Fishes," *Copeia*, pp. 507–540 (1967), created a new suborder Osmeroidei and modified the classification of GREENWOOD *et al.* (above); R.M. MCDOWALL, "Relationships of Galaxioid Fishes with A Further Discussion of Salmoniform Classification," *Copeia*, pp. 796–824 (1969), suggested further modifications in the classification of Salmoniformes; D.E. ROSEN and C. PATTERSON, "The Structure and Relationships of the Paracanthopterygian Fishes," *Bull. Am. Mus. Nat. Hist.*, 141:357–474 (1969), a revision of Salmoniformes with new information on early teleostean evolution; C. PATTERSON, "Two Upper Cretaceous Salmoniform Fishes from the Lebanon," *Bull. Br. Mus. Nat. Hist., Geol.*, 19:207–296 (1970), provides new information and suggested relationships of primitive salmoniforms; W.A. GOSLINE, "The Morphology and Systematic Position of the Alepocephaloid Fishes," *Bull. Br. Mus. Nat. Hist., Zool.*, 18:183–218 (1969), a review of suborder Alepocephaloidei; J.G. NIELSEN and V. LARSEN, "Synopsis of the Bathylaconidae (Pisces, Isospondyli) with a New Eastern Pacific Species," *Galathea Rep.*, 9:221–238 (1968), revises the suborder Bathylaconoidei, family Bathylaconidae in suborder Alepocephaloidei; N.B. MARSHALL, "*Bathyorion danae*, a New Genus and Species of Alepocephaliform Fishes," *Dana Rep.*, 68:1–10 (1966), a technical paper on the suborder Alepocephaloidei; G.J. NELSON, "Gill Arches of Some Teleostean Fishes of the Families Salangidae and Argentinidae," *Jap. J. Ichthyol.*, 17:61–66 (1970), another technical article revising the order Salmoniformes; R.J. BEHNKE, "A New Subgenus and Species of Trout, *Salmo (Platysalmo) platycephalus*, from Subcentral Turkey, with Comments on the Classification of the Subfamily Salmoninae," *Mitt. Hamb. Zool. Mus. Inst.*, 66:1–15 (1968), classification of trouts and salmons, and "The Application of Cytogenic and Biochemical Systematics to Phylogenetic Problems in the Family Salmonidae," *Trans. Am. Fish Soc.*, 99:237–248 (1970), classification of whitefishes, subfamily Coregoninae; STEPHEN D. SEDGWICK, *The Salmon Handbook: The Life and Cultivation of Fishes of the Salmon Family* (1982); GARY A. BORGER, *Naturals: A Guide to Food Organisms of the Trout* (1980).

Ostariophysi. JAMES W. ATZ, *Dean Bibliography of Fishes 1968* (1971), the first volume of a comprehensive, computerized bibliographic series; GEORGE ALBERT BOULENGER, *Catalogue of the Fresh-Water Fishes of Africa*, 4 vol. (1909-16), an important, illustrated, systematic account of Old World tropical groups; PHILIP J. DARLINGTON, JR., *Zoogeography: The Geographical Distribution of Animals* (1957), an excellent account of the distribution of freshwater fishes; CARL H. EIGENMANN and GEORGE S. MYERS (coauthor of pt. 5), *The American Characidae*, 5 pt. (1917–29), a classic treatise, only one-third completed upon death of the author; M.M. ELLIS, "The Gymnotid Eels of Tropical America," *Mem. Carneg. Mus.*, 6:109–204 (1914), a comprehensive, systematic, and morphological study; WILLIAM K. GREGORY and G. MILES CONRAD, "The Phylogeny of the Characin Fishes," *Zoologica*, 23:319–360 (1938), an old, somewhat equivocal, but important contribution on classification; HARRY GRUNDFEST, "Electric Fishes," *Scient. Am.*, 203:115–124 (1960), a semipopular but authoritative article; WILLIAM T. INNES, *Exotic Aquarium Fishes*, 19th ed. (1956), a well-illustrated, informative handbook of popular aquarium fishes; JOHN G. LUNDBERG and JONATHAN N. BASKIN, "The Caudal Skeleton of the Catfishes, Order Siluriformes," *Am. Mus. Novit. 2398* (1969), a description of anatomy, evolution, and relationships; W. PFEIFFER, "Alarm Substances," *Experientia*, 19:113–123 (1963), an excellent review article; C.T. REGAN, "The Classification of the Teleostean Fishes of the Order Ostariophysi," *Ann. Mag. Nat. Hist.*, Series 8, 8:13–32, 553–577 (1911), an old, but historically useful treatise on morphology and classification; JOHN H. TODD, "The Chemical Languages of Fishes," *Scient. Am.*, 244:99–108 (1971), a description of contemporary experiments on communication; STANLEY H. WEITZMAN, "The Osteology of *Brycon meeki*, a Generalized Characid Fish, with an Osteological Definition of the Family," *Stanford Ichthyol. Bull.*, 8:1–77 (1962), an important review of characid classification and osteology; JASPER S. LEE, *Commercial Catfish Farming*, 2nd ed. (1981).

Paracanthopterygii. *General:* C.M. BREDER and D.E. ROSEN, *Modes of Reproduction in Fishes* (1966); E.S. HERALD, *Living Fishes of the World* (1961).

Faunal: J.E. BÖHLKE and C.C.G. CHAPLIN, *Fishes of the Bahamas and Adjacent Tropical Waters* (1968); W.A. CLEMENS and G.V. WILBY, *Fishes of the Pacific Coast of Canada*, 2nd ed. (1961); A.H. LEIM and W.B. SCOTT, *Fishes of the Atlantic Coast of Canada* (1966); T.C. MARSHALL, *Fishes of the Great Barrier Reef and Coastal Waters of Queensland* (1964); Y. OKADA, *Fishes of Japan*, rev. ed. (1965); T.D. SCOTT, *The Marine and Freshwater Fishes of South Australia* (1962); J.L.B. SMITH, *The Sea Fishes of Southern Africa*, 4th ed. (1961); M. WEBER and L.E. DE BEAUEORT, *The Fishes of the Indo-Australian Archipelago* (1911–62), a multivolume work; A. WHEELER, *The Fishes of the British Isles and North-West Europe* (1969).

Specific: (*Batrachoidiformes*): B.B. COLLETTE, "A Review of the Venomous Toadfishes, Subfamily Thalassophryninae," *Copeia*, pp. 846–864 (1966); C.R. GILBERT, "Western Atlantic Batrachoidid Fishes of the Genus *Porichthys*, Including Three New Species," *Bull. Mar. Sci.*, 18:671–730 (1968). (*Gadiformes*): D.C. ARNOLD, "A Systematic Revision of the Fishes of the Teleost Family Carapidae (Percomorphi, Blennioidea), with Descriptions of Two New Species," *Bull. Br. Mus. Nat. Hist. (Zool.)*, 4:245–307 (1956); U. D'ANCONA and G. CAVINATO, *The Fishes of the Family Bregmacerotidae* (*Dana Rep.* 64) (1965); N.B. MARSHALL, "Systematic and Biological Studies of the Macrourid Fishes (Anacanthini-Teleosteii)," *Deep Sea Res.*, 12:299–322 (1965); J.G. NIELSEN, "Systematics and Biology of the Aphyonidae (Pisces, Ophidioidea)," *Galathea Rep.*, 10:1–90 (1969); D.W. STRASBURG, "Description of the Larva and Familial Relationships of the Fish *Snyderidia canina*," *Copeia*, pp. 20–24 (1965); A.N. SVETOVIDOV, *Gadiformes* (1962; originally published in Russian, 1948). (*Gobiesociformes*): J.C. BRIGGS, "A Monograph of the Clingfishes (Order Xenopterygii)," *Stanford Ichthyol. Bull.*, 6:1–224 (1955). (*Lophiiformes*): E. BERTELSEN, *The Ceratioid Fishes* (*Dana Rep.* 39) (1951); M.G. BRADBURY, "The Genera of Batfishes," *Copeia*, pp. 399–422 (1967); L.P. SCHULTZ, "The Frogfishes of the Family Antennariidae," *Proc. U.S. Natn. Mus.*, 107:47–105 (1957). (*Percopsiformes*): M.B. TRAUTMAN, *The Fishes of Ohio, with Illustrated Keys* (1957); L.P. WOODS and R.E. INGER, "The Cave, Spring, and Swamp Fishes of the Family Amblyopsidae of Central and Eastern United States," *Am. Midl. Nat.*, 58:232–256 (1957). *(Polymixiiformes)*: E.A. LACHNER, "Populations of the Berycoid Fish Family Polymixiidae," *Proc. U.S. Natn. Mus.*, 105:189–206 (1955).

Systematic: W.A. GOSLINE, *Functional Morphology and Classification of Teleostean Fishes* (1971); D.E. ROSEN and C. PATTERSON, "The Structure and Relationships of the Paracanthopterygian Fishes," *Bull. Am. Mus. Nat. Hist.*, 141:357–474 (1969).

Atheriniformes. R.M. ALEXANDER, "Mechanisms of the Jaws of Some Atheriniform Fish," *J. Zool.*, 151:233–255 (1967), an account of methods of jaw protrusion; C.M. BREDER and D.E. ROSEN, *Modes of Reproduction in Fishes* (1966), especially good on reproductive modifications in atheriniforms, with full bibliography; D.S. JORDAN and C.L. HUBBS, "A Monographic Review of the Family Atherinidae or Silversides," *Stanford Univ. Publs., Univ. Ser., Studies in Ichthyology*, 1:1–87 (1919), a classic review of atherinoids; C.T. REGAN, "On the Anatomy, Classification, and Systematic Position of the Teleostean Fishes of the Suborder Allotriognathi," *Proc. Zool. Soc. Lond.*, pp. 634–643 (1907), a classic paper in which the lampridiforms were first grouped together; R.R. ROEEN, "The Whale-Fishes: Families Cetomimidae, Barbourisiidae and Rondeletiidae (order Cetunculi)," *Galathea Rep.*, 1:255–260 (1959), an illustrated account of these deep-sea forms; D.E. ROSEN, "The Relationships and Taxonomic Position of the Halfbeaks, Killifishes, Silversides and their Relatives," *Bull. Am. Mus. Nat. Hist.*, 127:219–267 (1964), a monograph, with bibliography, in which the Atheriniformes were first grouped together, and with R.M. BAILEY, "The Poeciliid Fishes (Cyprinodontiformes), their Structure, Zoogeography, and Systematics," *ibid.*, 126:1–176 (1963), a monographic account of peociliids, and with C. PATTERSON, "The Structure and Relationships of the Paracanthopterygian Fishes," *ibid.*, 141:359–474 (1969), discussions of the relationships of fossil and living beryciforms, lampridiforms and atheriniforms, with bibliography.

Gasterosteiformes. AMERICAN EISHERIES SOCIETY, *A List of Common and Scientific Names of Fishes from the United States and Canada*, 3rd ed. (1970); LEO BERG, *Classification of Fishes Both Recent and Fossil* (1947), English and Russian; W.A. CLEMENS and G.V. WILBY, *Fishes of the Pacific Coast of Canada*, 2nd ed. (1961), limited in breadth and depth due to few order representatives in that range; EARL S. HERALD, *Living Fishes of the World* (1961); DAVID STARR JORDAN, *Fishes*, rev. ed. (1925), an excellent systematic work, although the nomenclature requires revision (college level); A.H. LEIM and W.B. SCOTT, *Fishes of the Atlantic Coast of Canada* (1966), a work broader in scope than Clemens-Wilby (above); N. TINBERGEN, "The Curious Behavior of the Stickleback," *Scient. Am.*, 187:22–26 (1952), a popularized segment of the author's 1951 *Study of Instinct;* GILBERT WHITLEY and JOYCE ALLAN, *The Sea-Horse and Its Relatives* (1958), a brief, supposedly popularized account of order members in Australian waters—of amazing scope, interest, and depth.

Scorpaeniformes. Most of the information on the scorpaeniform fishes is found in short technical articles in scientific journals. The following works are broad in nature but contain substantial information on members of this order. E.S. HERALD, *Living Fishes of the World* (1961), is a well-illustrated book for the general reader; N.B. MARSHALL, *The Life of Fishes* (1965), contains some sections on the biology of scorpaeniforms; G.V. NIKOLSKII, *Special Ichthyology*, 2nd rev. ed. (1961; trans. of the 2nd Russian ed. of 1954), deals especially with species commercially important in the Soviet Union; B.W. HALSTEAD, *Poisonous and Venomous Marine Animals of the World*, vol. 3 (1970), provides information on the venom of these fishes and its effects. P.H. GREENWOOD *et al.*, "Phyletic Studies of Teleostean Fishes, with a Provisional Classification of Living Forms," *Bull. Am. Mus. Nat. Hist.*, 131: 339–455 (1966); and C.T. REGAN, "The Osteology and Classification of the Teleostean Fishes of the Order Scleroparei," *Ann. Mag. Nat. Hist.*, Series 8, 11:169–184 (1913), are rather technical, but important, contributions to the study of the scorpaeniforms.

Perciformes. Recent books containing many excellent illustrations of perciform fishes and short accounts of the biology of each group are: J.E. BÖHLKE and C.C.G. CHAPLIN, *Fishes of the Bahamas and Adjacent Tropical Waters* (1968), a well-illustrated work with keys for identifying species and excellent summaries of the biology of each family; DAVID STARR JORDAN and BARTON WARREN EVERMANN, *The Fishes of North and Middle America*, 4 vol. (1896–1900, reprinted 1963); T.C. MARSHALL, *Fishes of the Great Barrier Reef and Coastal Waters of Queensland* (1964), contains many colour and black-and-white pictures of perciform fishes; I.S.R. MUNRO, *The Fishes of New Guinea* (1967), 1,095 fishes illustrated by photographs, with 76 species in colour; J.L.B. SMITH, *The Sea Fishes of Southern Africa*, 4th ed. (1961), profusely illustrated with many colour plates; J.L.B. and MARGARET M. SMITH, *The Fishes of Seychelles* (1963), 880 species illustrated, many in colour; A.C. WHEELER, *The Fishes of the British Isles and North-West Europe* (1969); JOHN E. RANDALL, *Caribbean Reef Fishes* (1968), contains many fine photographs of perciforms and others; ROBERT B. CHIASSON, *Laboratory Anatomy of the Perch*, 3rd ed. (1980).

Other books with information about perciform fishes are EARL S. HERALD, *Living Fishes of the World* (1961); N.B. MARSHALL, *The Life of Fishes* (1966); and U. OKADA, *Fishes of Japan* (1955).

References pertaining to classification and relationships include: W.C. EREIHOEER, "Patterns of the Ramus Lateralis Accessorius and Their Systematic Significance in Teleostean Fishes," *Stanford Ichthyol. Bull.*, 8:79–189 (1963); and "Trunk Lateral Line Nerves, Hyoid Arch Gill Rakers, and Olfactory Bulb Location in Atheriniform, Mugilid, and Percoid Fishes," *Occ. Pap. Calif. Acad. Sci.*, no. 95 (1972); W.A. GOSLINE, "Systematic Position and Relationships of the Percesocine Fishes," *Pacif. Sci.*, 16:207–217 (1962); "The Suborders of Perciform Fishes," *Proc. U.S. Natn. Mus.*, 124:1–78 (1968); and *Functional Morphology and the Classification of Teleostean Fishes* (1971); P.H. GREENWOOD *et al.*, "Phyletic Studies of Teleostean Fishes, with a Provisional Classification of Living Forms," *Bull. Am. Mus. Nat. Hist.*, 131:339–455 (1966); D.E. MCALLISTER, "The Evolution of Branchiostegals, and Associated Opercular, Gular, and Hyoid Bones, and the Classification of Teleostome Fishes, Living and Fossil," *Bull. Natn. Mus. Can.*, no. 221 (1968); C.T. REGAN, a series of papers on perciform classification in *Annals and Magazine of Natural History*, series 7, vol. 11 and series 9, vol. 11 (1903–23); and "Fishes," *Encyclopaedia Britannica*, 14th ed., vol. 9, pp. 305–328 (1929), which still forms the major basis of present perciform classification; D.E. ROSEN, "The Relationships and Taxonomic Position of the Halfbeaks, Killifishes, Silversides, and Their Relatives," *Bull. Am. Mus. Nat. Hist.*, 127:217–267 (1964), and with COLIN PATTERSON, "The Structure and Relationships of the Paracanthopterygian Fishes," *Bull. Am. Mus. Nat. Hist.*, 141:357–474 (1969).

Pleuronectiformes. J.T. CUNNINGHAM, *A Treatise on the Common Sole,* Solea vulgaris, *Considered Both as an Organism and as a Commodity* (1890), a monograph on the morphology, development, life history, and economics of this species, including a taxonomic review; C.L. HUBBS, "Phylogenetic Position of the Citharidae, a Family of Flatfish," *Misc. Publs. Mus. Zool. Univ. Mich.*, no. 63, pp. 5–38 (1945), presents a definition of the family Citharidae and phylogenetic relationships of the flatfish families; H.M. KYLE, "Flat-fishes (Heterosomata)," *Rep. Danish Exped. Medit.* 2:1–150 (1913), a major treatise on the development of some flatfishes in the families Bothidae, Soleidae, and Cynoglossidae (illustrations of larvae included when known); and "The Asymmetry, Metamorphosis and Origin of Flatfishes," *Phil. Trans. R. Soc.*, Series B, 211:75–129 (1921), a major paper on the origins of pleuronectiform metamorphosis; J.R. NORMAN, *A Systematic Monograph of the Flatfishes (Heterosomata)*, vol. 1, *Psettodidae, Bothidae, Pleuronectidae* (1934), a classic on the taxonomy of flatfishes, the only paper dealing with these families on a worldwide basis; B. BENNET RAE, *The Lemon Sole* (1978).

Tetraodontiformes. J.E. BÖHLKE and C.C.G. CHAPLIN, *Fishes of the Bahamas and Adjacent Tropical Waters* (1968), includes an excellent combination of scientific and popular accounts of the Caribbean tetraodontiforms, with all of the species well illustrated. BRUCE W. HALSTEAD, *Poisonous and Venomous Marine Animals of the World*, vol. 2, *Vertebrates*, rev. ed. (1978), contains a comprehensive review of the poisonous properties of the tetraodontiforms, with numerous illustrations of poisonous species. JAMES C. TYLER, *A Monograph on Plectognath Fishes of the Superfamily Triacanthoidea* (1968), is a technical monograph on the two most generalized families of tetraodontiforms but also includes general accounts of the way of life, the distribution, and the relationships of these families, as well as an extensive bibliography on related articles.

Commercial Fishing

Fishing, which involves the recovery of food and other valuable resources from bodies of water, is one of the oldest employments of mankind. Ancient heaps of discarded mollusk shells, some from prehistoric times, have been found in coastal areas throughout the world, including those of China, Japan, Peru, Brazil, Portugal, and Denmark. These mounds, known as kitchen middens (from the Danish *køkkenmødding*), indicate that marine mollusks were among the early foods of humans.

Archaeological evidence shows that humans next learned to catch fishes in traps and nets. These ventures were limited at first to the lakes and rivers, but as boats and fishing devices were improved, humans ventured into sheltered coastal areas and river mouths and eventually farther out onto the continental shelves, the relatively shallow ocean plains between the land and the deeper ocean areas. In some shelf areas where seaweed was abundant, this was also incorporated into the diet.

Fishing technology continued to develop throughout history, employing improved and larger ships, more sophisticated fishing equipment, and various food preservation methods. Commercial fishing is now carried on in all types of waters, in all parts of the world, except where impeded by depth or dangerous currents or prohibited by law. Commercial fishing can be done in a simple manner with small vessels, little technical equipment, and little or no mechanization as in small local, traditional, or artisanal fisheries. It can also be done on a large scale with powerful deep-sea vessels and sophisticated mechanical equipment similar to that of modern industrial enterprises.

Both plants and animals are taken from the sea. Two types of fish are caught: demersal, living at or near the bottom, although sometimes in mid-water; and pelagic, living in the open sea near the surface. Cod, haddock, hake, pollock, and all forms of flatfish are common demersal fish. Herring and related species and tuna and their relatives are examples of pelagic fish. Both demersal and pelagic fish can sometimes be found far from coastal regions. Other aquatic animals that may be the object of commercial fishery include, most notably, crustaceans (lobsters, spiny lobsters, crabs, prawns, shrimps, crayfish) and mollusks (oysters, scallops, mussels, snails, squid, octopuses). Certain mammals (whales, porpoises), reptiles (serpents, crocodiles), amphibians (frogs), many types of worms, coelenterates (coral, jellyfish), and sponges are also sought by commercial fishermen. Most of these animals are legally regarded as fish in many countries.

The most important water plants commercially obtained in seawater and fresh water are algae. Seaweed is harvested in the water or collected on the seashore. Algae play an important ecological role in many countries, not only as human food but also as fodder for cattle, as fertilizer, and as a raw material for certain industries.

Fisheries are classified in part by type of water: fresh water—lake, river, and pond—and salt water—inshore, mid-water, and deep sea. Another classification is based on the object—as in whaling, salmon fishing, and sponge fishing. Sometimes fisheries are classified according to the method of fishing employed: harpooning, seining, trawling, and lining.

Whaling, historical and modern, has been documented in records from Neolithic cave art to present-day annual reports of the International Commission on Whaling, with no firm proof as to what people first engaged in this dangerous activity. Lacking adequate agriculture, early inhabitants of the polar littoral developed successful whaling techniques using Stone Age weapons. When first contacted, the Inuit (Eskimos) of eastern and western North America showed mastery in whale hunting, and much the same methods were in use up to 1900. For them a captured whale supplied food, fuel, and light; sinews provided cordage, and skeletal parts were used for tools and construction. Not until the 20th century, when floating factories came into use, did another civilization succeed in the same efficient use of the whole carcass. A possible exception is Japan, where whale flesh has been popular from early times. Elsewhere, however, from the first intensive exploitation of whales in the early 17th century to quite recent times, little more than blubber was used, the remainder being discarded. Each successive discovery of a new whaling ground resulted in the near disappearance of a particular species. The efficiency of modern hunting methods speeded this trend to the point that the industry has all but taken its place in history, leaving only a few enterprises to carry on in a limited fashion.

This article discusses organized fishing for profit, with an emphasis on mechanized industrial methods, gear, and vessels. The history and methods of whaling, which is less fishing than the hunting of an aquatic mammal, are discussed separately in this article from freshwater and saltwater fishery and aquaculture. For angling, or recreational fishing, see the *Micropædia* article FISHING. For information on the use and value of fish and marine products as food, see the *Macropædia* article NUTRITION.

(A.v.B./G.A.B./P.F.P./Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 731.

This article is divided into the following sections:

History of commercial fishing

Food-gathering people first obtained fish and shellfish from the shallow water of lakes and along the seashore, from small ponds remaining in inundation areas, from tidal areas, and from small streams. Some authorities believe that in the earliest times fish were rarely caught because of the inadequacy of fishing gear. Shellfish, however, can be gathered easily by hand, and the prehistoric kitchen middens indicate their importance as a food source.

In earliest times most foodstuffs were used at once and not stored, but as expanding populations increased food needs, techniques were developed for preserving fish by drying, smoking, salting, and fermentation. It became desirable to catch large quantities, and specialized equipment was devised. Individual fishing was replaced by collective efforts involving larger, more effective gear.

Fishing equipment and methods improved through the centuries, until bulk fisheries were established in Europe. Herring were caught in huge numbers in northern Europe in the Middle Ages. Cod fishery began on the Grand Banks of Newfoundland even before the Italian explorer John Cabot made his voyage there in 1497. Whaling with large fleets began in the 17th century, both in the Atlantic and in the South Pacific.

Fishing under sail

Before mechanization came to the fishing industry toward the end of the 19th century, sailing vessels developed to suit conditions and fisheries in different areas. The Grand Banks schooners were the peak of such developments. Sailing from New England, Nova Scotia, and Newfoundland, they fished cod during trips lasting up to six months, salting the catch for export to Europe, Africa, and the Caribbean. Individual fishermen fished from small wooden dories, setting and hauling longlines by hand. Portuguese vessels also sailed annually to the Grand Banks, and a number still operated alongside modern steel vessels into the late 20th century. Smaller cutters and yawls worked around Europe using drift gill nets and setnets. Beam trawls were used extensively in the North Sea and English Channel, particularly for flatfish, being towed downwind under sail and then hauled back to the vessel's side.

As steam-driven winches came into use, fishing gear increased in size and weight. Steam gradually replaced sail for propulsion in the last quarter of the 19th century. In turn, the internal-combustion engine supplanted steam, although steam trawlers continued operating as late as the 1950s. Smaller craft became motorized in the early 20th century, and the inboard diesel engine became universally adopted—except for the smallest boats, on which gasoline-powered outboard engines remain common.

Larger catches could be obtained by increasing the number or the size of the fishing gear or both. Simple lines armed with one or a few hooks were replaced by longlines with thousands of hooks. Single small traps were combined into systems of hundreds, and pots were set in large quantities. Nets were greatly enlarged; netmaking machines were invented that produced netting in large sheets. Mechanical netmaking brought replacement of the old local netting fibres (linen and hemp) with cotton and hard fibres. But all natural fibres, especially those of cellulose, begin to rot in time; thus, the introduction after World War II of rot-proof nets made of synthetic fibres represented a major advance. Mechanical netmaking remained unchanged for the most part, though for certain fishing gear the usual knotted netting was replaced by knotless netting.

Mechanized net handling

In the beginning of the 1950s, mechanization took a great stride forward in purse seining when the power block was invented for hauling the gear. Another important hauling device was a power-driven drum to haul and store seine nets, gill nets, purse seines, and even the large trawl nets. The Japanese introduced drums in longline fishing for tuna. Another important innovation was the stern chute for stern trawlers, a development made possible by cooperation between naval architects and fishing-gear experts, which permitted large-scale mechanization of gear handling.

An era of rapid technical development in vessel design began with the British factory trawler experiment in the late 1940s, which demonstrated the great advantage of large stern trawlers that processed their catch on board. The idea was quickly developed by countries seeking to fish distant resources, and by the mid-1960s these large vessels (up to 100 metres long) were operated by the Soviet Union, the United Kingdom, Japan, Poland, East Germany, and Spain.

Equivalent development occurred in the exploitation of the huge resources of small pelagic fish, mainly for conversion into fish meal. In the late 1940s small vessels, using hand-operated natural-fibre nets, fed small shore-based canning and fish-meal plants. By the late 1960s, large fleets of 25-metre purse seiners were supplying factory mother ships capable of handling up to 3,000 tons per day.

Concurrently, developing countries strove to introduce more modern fishing technology in order to boost protein supplies for their populations. Most rely heavily on artisanal fishing, using canoes or small boats with simple gear and often working off of open beaches. The introduction of outboard motors, larger boats, and synthetic nets enabled many countries to increase their catches significantly.

In the 40 years following World War II, the annual world fishing catch quadrupled. By the early 1970s, though, it had become apparent that such development was not limitless. Several of the largest resources of pelagic fish harvested by purse seiners suffered collapses generally blamed on overfishing. These included the northeast Atlantic herring, the South Atlantic pilchard, and the West African sardine and associated species. Severe declines in catches of stocks fished by fleets of factory trawlers caused such concern that coastal states pressed for protection of the resources off their shores. In 1972 Iceland became the first country to claim an extended fisheries limit of 50 miles (93 kilometres) and, in 1975, 200 miles. Other countries followed suit, and in 1983 the Law of the Sea established an exclusive economic zone, or EEZ, of 200 miles, inside of which each country had exclusive right to the exploitation of marine life. An immediate result was the exclusion from many areas of high-performance, long-distance foreign fleets, which were replaced by often less-efficient domestic coastal craft. For example, the British fleet of 168 distant-water trawlers disappeared within a few years, to be replaced by a fleet of compact, coastal-type vessels.

Rising fue costs

The oil crisis of the 1970s increased fuel costs as much as 400 percent, while fish prices rose by only about 80 percent. This forced many fuel-inefficient vessels, such as many of the U.S. Gulf shrimp trawlers, to tie up or transfer to other fisheries. Resulting development of fuel-efficient vessels, engines, fishing methods, and equipment—including applications of modern sail technology—depended thereafter upon the price of oil.

With the growing importance of managing fisheries to ensure maximum possible benefit from a particular stock, the work of fisheries scientists increased in importance. From a mainly descriptive science in the 19th century, the field evolved, especially after World War II, to develop sophisticated computer analyses based on mathematical models to predict the optimum yields available from fish populations. (For a history of the whaling industry, see below *Whaling.*)

Fishery equipment and facilities

GEAR

An international classification of fishing methods includes 16 categories, depending upon the fishing gear and the manner in which the gear is used: (1) fishing without gear, (2) grappling and wounding gear, (3) stunning, (4) line fishing, (5) trapping, (6) trapping in the air, (7) fishing with bag nets, (8) dredging and trawling, (9) seining, (10) fishing with surrounding nets, (11) driving fish into nets, (12) fishing with lift nets, (13) fishing with falling gear, (14) gillnetting, (15) fishing with entangling nets, and (16) harvesting with machines.

Hand tools. The simplest and oldest form of fishing, collecting by hand, is still done today by both professionals and nonprofessionals along the shore during ebb tide in shallow water and in deeper water by divers with or without diving suits. Even when small tools such as knives or hoes are used, such collecting is classified as without gear. Diving to collect sponges, pearl oysters, or corals belongs under this classification, as does fishing with hunting animals. The Chinese still use trained otters, and the Japanese sometimes employ cormorants.

To extend the reach of the human arm, long-handled tools were invented, such as spears, which can be thrust, thrown, or discharged, and clamps, tongs, and raking devices for shellfish harvesting. A special form is the harpoon, composed of a point and a stick joined together by a rope. Such grappling and wounding gear also includes spears, blowpipes, bows and arrows, and rifles and guns, which are used in fish shooting.

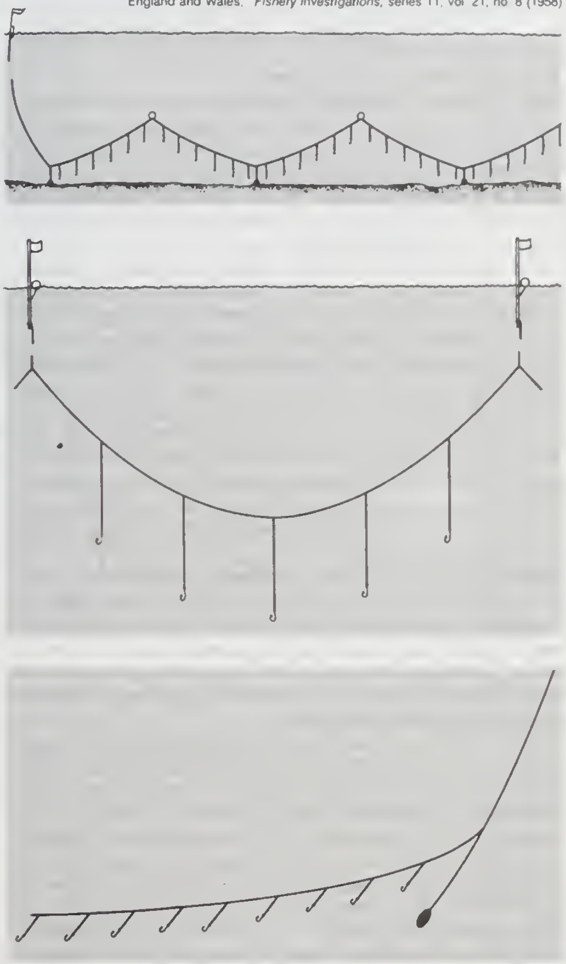
The method called stunning may involve poisoning with toxic plants and special chemicals or mechanical stunning by explosions under water. The most modern practice in this field is to stun the fish by means of an electrical shock.

Lines. In line fishing the fish can be attracted by a natural or artificial bait or lure devised to catch and hold the fish. Generally, the bait is combined with a hook or with a gorge, as is used in France in line fishing for eels. There are handlines, as in pole-and-line fishing for tuna; setlines, such as bottom longlines with hundreds of hooks, used for cod or halibut; drift lines with a single hook and drifting longlines for tuna, shark, and salmon; and troll lines for mackerel and some game fish. Another method of fishing with hooks is done without bait, by raising and lowering arrays of hooks to gig (hook in the body) such large species as cod and sturgeon.

Handlines,
longlines,
and troll
lines

Traps. Genuine mechanical traps, which close by a mechanism released by the prey, are seldom employed for fishing. Most commercial fishing traps are chambers entered easily by the prey but from which escape is prevented by labyrinths or retarding devices, such as gorges or funnels. Fish traps can be simple hiding places, such as bushes or tubes, into which fish or shrimps swim for shelter but cannot escape later when the device is hauled in. The octopus pot used on the Italian coast and by the fishermen of South and East Asia is an example. Other types include small basketlike or cagelike traps made of wood, netting, wire, or plastic pots and fyke nets (long bag-shaped nets kept open by a series of hoops). Large pound nets, composed of net walls that guide fish through a series of baffles into a catching area, are used in the Mediterranean for tuna, off the western Baltic coast for eels, herring, and other species, and off both coasts of the northern Pacific for salmon. A special class are aerial traps for catching flying fish and shrimps. The fish are stirred up, then caught in the air with the help of special gear called veranda nets. South Sea islanders catch flying fish at night by attracting them with torches.

From A. von Brandt, *Fish Catching Methods of the World*, 3rd ed. (1984), Fishing News Books Ltd., after (top) N. Peters, *Handbuch der Seefischerei Nordeuropas* 4 (1935), and (bottom) F.M. Davis "An Account of Fishing Gear of England and Wales," *Fishery Investigations*, series 11, vol. 21, no. 8 (1958)



(Top) Bottom longline, (centre) drifting longline, (bottom) troll line.

Nets. *Bag nets.* Bag nets are kept vertically open by a frame and held horizontally stretched by the water current. There are small scoop nets that can be pushed and dragged and big stownets, with and without wings, held on stakes or on anchors with or without a vessel. There is also a special winged type with boards or metal plates (called otter boards) that keep it spread open. Stownets, larger than scoop nets and held in place against a current, are used in many rivers and by the Koreans for sea fishing in the strong current off the southwest coast of their country. In this case the stownet is anchored with a vessel.

Dragged nets. Dragged gear includes dredges, which are used mostly for shellfish and may be operated by hand in shallow waters or towed from large vessels. Another dragged net is the trawl, a large, cone-shaped bag of netting that is dragged along the seabed or towed in mid-water between the seabed and surface. Trawls are the most important fishing gear of the commercial fisheries of northwest Europe and are second only to purse seines in total catch of the world.

Seines. The seine net has very long wings and towing warps (tow lines), with or without bags for the catch. With purse seines, pelagic fish are surrounded not only from the side but also from underneath, preventing them from escaping by diving downward. Purse seines can be operated by a single boat, with or without auxiliary skiff, or by two vessels. Many sardinelike fishes—herring, tuna, mackerel, cod, and salmon—are commercially fished in this manner.

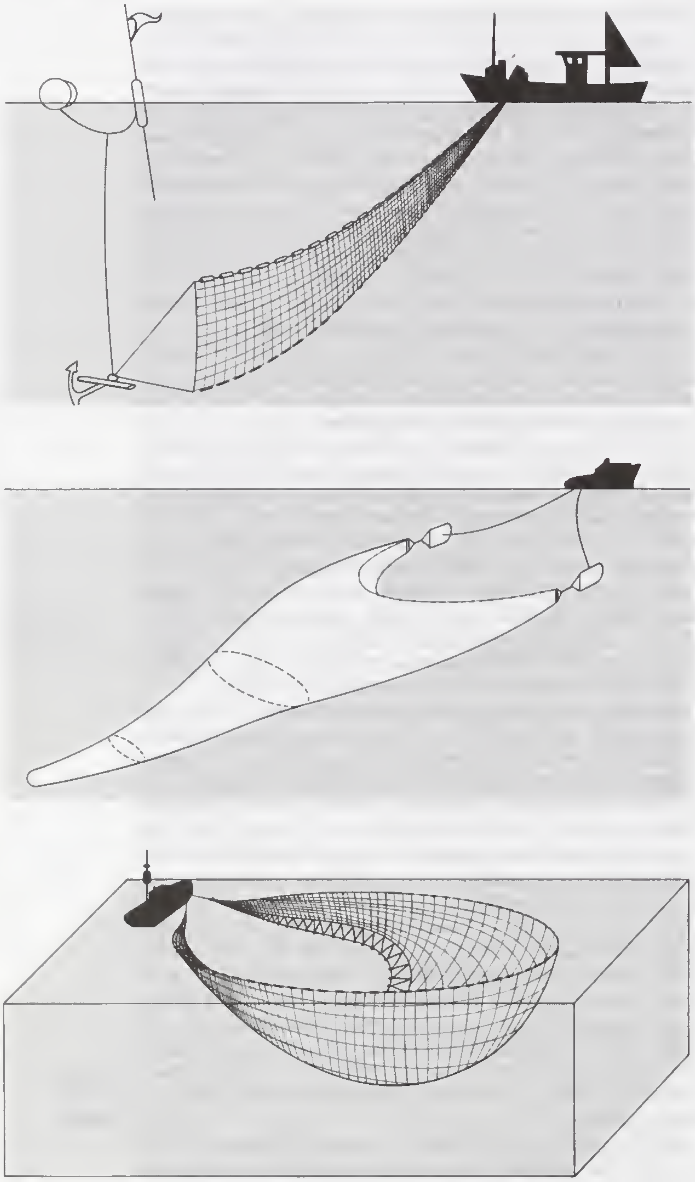
Objects
of seine
fishery

Drive-in and lift nets. Another class of fishing methods involves driving the fish into a net or gear. A drive-in net may be one of those already mentioned or may be specially made, such as the dustpan-shaped stationary gear used in some fisheries in South Asia.

A further fishing method employs lift nets, which are submerged, then raised or hauled upward out of the water to catch the fish or crustaceans above them, often attracted by light or natural bait. This group includes small hand-operated lift nets, such as hoop and blanket nets, as well as large, mechanically and pneumatically operated lift nets. Some of these employ levers, or gallows, and are installed on the beach or on a vessel. The fish wheels used on the Tiber, Rhône, and Columbia rivers can be considered as mechanized lift nets. The most important examples of this fishing method are the stick-held dip nets of the Japanese. In contrast to the lift nets are falling gear, which can be wooden baskets, cover pots, or a variety of nets designed to be cast on fish and crustaceans from above.

Gill and entangling nets. Gill nets, which catch the fish in their meshes, are mostly used in long rows. As setnets they are anchored or fixed by stakes; as drift nets they drift freely or with a fishing craft. Before the invention of mid-water trawls, drift nets, with surrounding nets, were the principal gear for fishing pelagic fishes.

Sometimes gill nets do not catch by meshing but by entangling the fish, especially those too large for the mesh size or provided with spines or hard fins. Single-walled tangle nets are widely used to catch sturgeon, salmon, and shellfish, such as the king crab. Some tangle nets are double walled; most are triple walled, such as the trammel nets used especially for flatfish.



(Top) Gill net, (centre) trawl, (bottom) purse seine.

(Top and centre) By courtesy of the Fish and Wildlife Service, U S Department of the Interior, (bottom) from A von Brandt, *Fish Catching Methods of the World*, 3rd ed (1984), Fishing News Books Ltd , after U Bertuccioli, *Il primo libro del pescatore*, Venice (1955)

Machines. Harvesting machines include comparatively new types of gear that may separate the fish or shellfish from the water by pumps (pump fishing) or by mechanized dredges, as well as floating machines that dig out mollusks by means of underwater jets and transport them out of the water with the help of conveyor belts.

(A.v.B./J.C.Sa.)

VESSELS

Until the mid-20th century, fishing boats were largely of local design, with different types found even in adjacent ports. As fishermen started to roam farther afield for their catches, the vessels grew, and with this growth in size came an element of standardization in design. Today, fishing boat design and construction is an international industry, with the different vessel types dictated more by the fishing methods for which they are designed rather than by their port or country of origin.

The establishment of 200-mile fishing limits (see above *History of commercial fishing*) has altered fishing patterns and, with them, the types of vessels used by many countries. In the United States and Canada, fishing vessels have grown with the introduction of processing or factory trawlers, while the huge fleets of this type of vessel operated by Soviet-bloc countries and Japan have shrunk. In western Europe, compact fishing vessels have been developed with high catching power. The advantage of these smaller vessels is their reduced capital and operating costs.

Construction materials

Steel is the most common construction material, being used exclusively on larger vessels (above 25 metres). Traditional wood construction is less common because of cost and a lack of suitable timber in many areas. The use of fibreglass is increasing in smaller fishing vessels, and it is now used on vessels of up to 25 metres in length. Ferrocement has been used to a certain extent; it is mainly used in the artisanal fisheries of developing countries because, while its construction is labour intensive, its raw materials are cheap.

The aim in all fishing-boat development is to improve efficiency by building vessels that have higher catching power, smaller crews, and reduced operating costs. This development must be matched against safety concerns, as commercial fishing is one of the highest risk industries in the world. Several countries have introduced regulations governing the construction and operation of fishing vessels. The International Maritime Organization, convened in 1959 under the auspices of the United Nations, is responsible for devising international regulations covering such aspects of fishing vessel design as construction, stability, safety equipment, and watertight integrity. These regulations are likely to lead to further standardization in design.

The Food and Agriculture Organization of the United Nations has introduced a classification scheme of fishing vessels based primarily on the gear used.

Trawlers. Most trawlers are single-screw vessels with powerful engines and deck machinery for dragging the trawl nets.

Side trawlers. On this traditional type of trawler, the trawl is launched and recovered from the side of the vessel. The side trawler is characterized by the wheelhouse and superstructure at the stern and a raised forecastle at the bow. The hull lines follow a traditional seaworthy form, with a pronounced deck sheer giving a high bow and stern. The working deck may be covered.

Stern trawlers. Practically all trawlers built today are stern trawlers, with the trawl launched and recovered over the stern. The vessels are generally designed with the wheelhouse and superstructure forward, often forming part of the raised forecastle. By contrast, the working deck aft is lower, and, on the larger trawlers a ramp is built into the stern up which the trawl is pulled onto the deck. On smaller stern trawlers the trawl is lifted on board by a hoist.

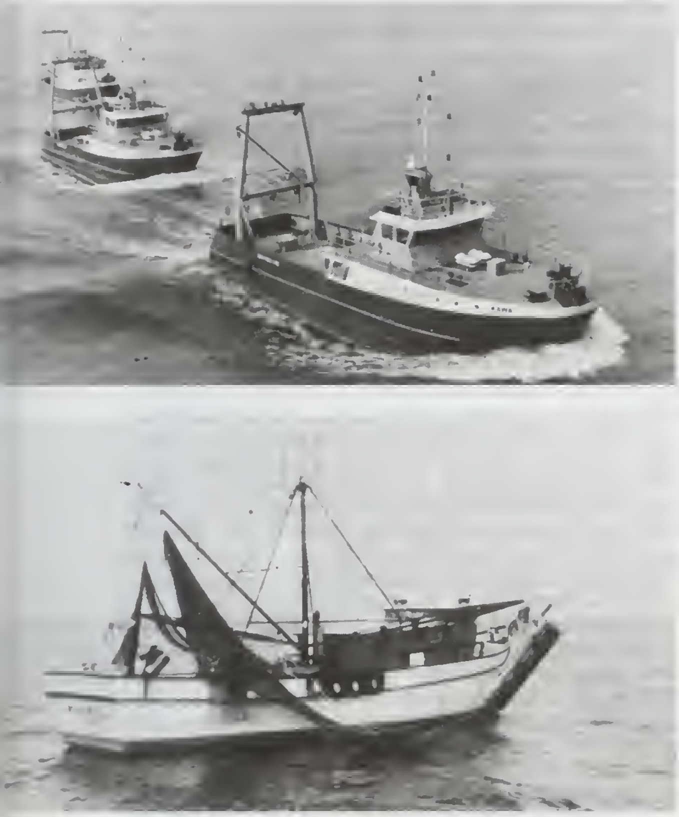
Beam or outrigger trawlers. With this type of vessel, two beam trawls are towed from booms extending to each side and supported by a central mast. The booms are very strong, as they take the full weight of the trawl being towed. The mast supporting the booms may be located forward, in which case the wheelhouse is located aft as on a side trawler, or they may be amidships with the wheelhouse forward, as on a stern trawler. The former type is widely used for beam trawling in Europe, while the latter is the pattern of most shrimp trawlers. European-style beam trawlers are the most powerful fishing vessels of their size in the world.

Shrimp trawlers

Wet-fish trawlers. This type is distinguished by the way the catch is stored on board. It can be either a side or stern trawler with an insulated hold where the fish are stored "wet," or fresh, after sorting. Ice used to cool the catch may be loaded at the start of the voyage or produced on board. This type of trawler normally operates on fishing trips lasting less than four days.

Freezer trawlers. On this type, constituting most large trawlers, the catch is preserved by freezing. On some vessels the catch is gutted and sorted before freezing, but processing is done mainly after the catch is landed.

Factory or processing trawlers. These are the largest type of fishing vessel. After catching and sorting, the fish is transferred to the processing deck, where it is processed and packaged. It is then frozen and stored in the hold. Many vessels have facilities for extracting oil and for making fish meal from waste products. Factory trawlers accommodate large crews and stay at sea for many weeks. They often support a fleet of smaller trawlers; when they



Two types of trawler.
(Top) Modern Japanese stern trawlers, with raised foredeck and low working afterdeck. (Bottom) American shrimp trawler, showing the booms typical of outrigger trawlers.
Dag Pike

load fish from other vessels rather than catching it themselves, they are called Klondykers.

Seiners. Seiners range in size from canoes, where the net is hauled by hand, to larger vessels with powerful net-handling equipment. This equipment generally consists of a power block mounted on a crane placed aft of the wheelhouse, as well as winches and drums for hauling and storing the great lengths of net and rope required for seine fishing.

Purse seiners. In purse seining, the fish shoal is surrounded by the net, which has a rope that seals the bottom of the net to trap the fish. Small fish may be pumped out of the net, or the net can be hauled on board and the fish released for sorting.

North American and European purse seiners

The North American purse seiner is generally laid out with a forward wheelhouse and aft working deck. A characteristic of this vessel is the powerful net block on the end of a long boom supported by the mast and a crow's nest on the mast for spotting fish shoals.

European purse seiners are generally larger than their North American counterparts, being typically 30 metres in length. They have an aft wheelhouse and the net is hauled at the stern. The herring and mackerel caught by these vessels, needing sensitive handling and storage, are stored in tanks of chilled seawater built amidships in the hull. Thrusters (screws that provide sideways movement) are usually fitted to these vessels to give improved maneuverability when laying and hauling nets.

Tuna purse seiners are large vessels mainly designed for long-range fishing, although smaller types operate in the Mediterranean. They are similar in design to, but larger than, the North American purse seiner, and they have a sloping stern where a tuna skiff, used for laying the net, is stowed. Several smaller boats are also carried to help with handling the catch and removing unwanted or protected fish from the nets. In addition to a crow's nest for spotting fish shoals, a light helicopter is sometimes carried on a helicopter deck above the forward wheelhouse. Modern tuna vessels store the catch in chilled seawater tanks.

Seine-netters. These vessels, like the European purse seiner, have the wheelhouse placed aft. Rather than purse seines, they employ nets similar to bottom trawls, which they set on long ropes and then haul in along the bottom like seine nets. Winches and reels on the forward deck haul and stow the ropes, while a power block aft hauls in the net.

Dredgers. These vessels tend to fish in sheltered and shallow waters for certain types of shellfish. They are similar to beam trawlers, but they may have four booms for towing the dredges. The hulls are often shallow-draft, and hand or mechanical sorting facilities are fitted on deck. They may have forward or aft wheelhouses.

Lift-netters. These vessels catch fish by lowering nets over the side, switching on powerful lights to attract the fish, and then lifting the net. Their main characteristics are long booms and support masts along the working side of the vessel. Lift-netters are generally low-powered vessels working on short trips.

Gill-netters. Gill nets are used by all sizes of fishing boat up to 20 metres in length. There is no characteristic style, although this type of vessel often uses a steadying sail to keep heading into the wind. The nets may be set and hauled by hand or by power blocks at deck level.

King crab potters

Potters. These are generally inshore vessels using pots or traps to catch shellfish. They come in a wide variety of types and sizes, but a typical inshore potter is 10 metres in length. King crab potters working off of the coast of Alaska are up to 30 metres in length; they generally have the wheelhouse forward to leave a clear working deck aft, but smaller vessels can have the wheelhouse at either end. A characteristic of potters is the pot stowage, which is usually a large frame construction aft. Gear is handled with a crane on larger vessels and with a bulwark-mounted pot hauler on smaller vessels.

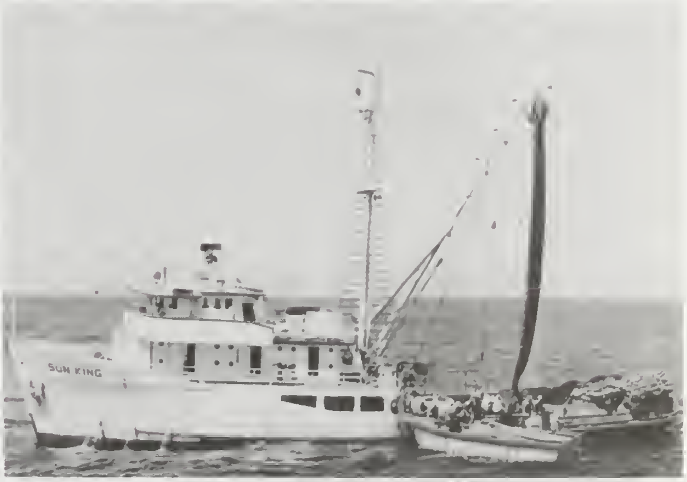
Liners. Fishing with lines and hooks is carried out by a wide range of vessels using either manual or mechanical hauling.

Handliners. These are generally small fishing boats, open or decked, working inshore waters.

Longliners. These tend to be larger vessels with the hooks and lines attached to a rope that is supported by floats or simply trailed. Usually there is an automatic line system whereby the hooks are baited and fish removed mechanically in what can be a continuous system. As line-caught fish tend to be of the best quality, chilled seawater tanks are often installed to maintain freshness. The largest types of longliner are those fishing for tuna; these can be more than 60 metres in length.

Pole-and-line vessels. These vessels use lines mounted on poles or fishing rods, hand operated on smaller vessels and mechanized on larger vessels. A large crew is required, and, typically, the vessels are built with the entire deck edge clear to give the maximum fishing area. The wheelhouse can be forward or aft.

By courtesy of the U.S. Department of the Interior, Fish and Wildlife Service



North American purse seiner, with wheelhouse forward and power block for working the net.

Small lobster boat loaded with metal traps, New Harbor, Maine.

E R Degginger/Bruce Coleman Inc

Trollers. On trollers the lines are trailed aft from twin booms. These vessels are similar to beam trawlers, but they have much lighter gear and almost invariably have the wheelhouse forward so that the lines can be hauled in on the working deck aft.

Multipurpose fishing boats. Because fishing for certain species of fish is often seasonal, many modern fishing boats are designed to incorporate two or more different fishing methods. Typical is the trawler/purse seiner, but potting vessels and longliners can also be equipped for trawling. Trawlers can also work at pair trawling, in which a trawl is pulled between two vessels. This may require heavier gear to handle the larger trawl.

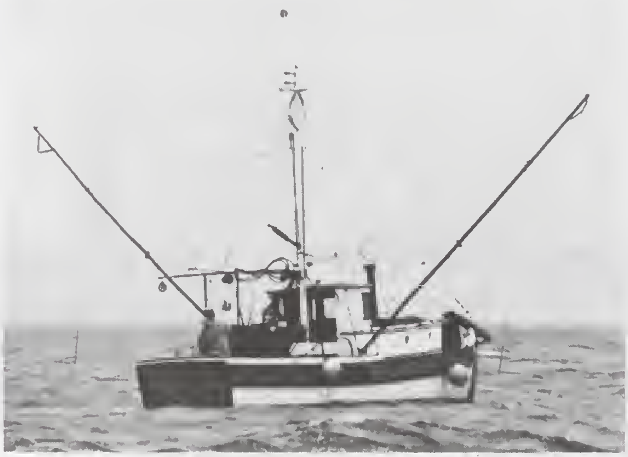
Artisanal fishing boats. These are generally used in less-developed countries, working off of an open beach and using very basic fishing methods with no mechanization. Sails and oars are often used as motive power, but under government aid programs there have been advances in design, and engines—mainly outboard motors—are becoming common. Many of the advanced designs retain traditional characteristics to make them acceptable to the fishermen.

Mother ships. This category generally covers vessels carrying small fishing boats that return to the mother ship with their catch. They are generally ocean-going vessels with extensive on-board facilities for processing and freezing the catch. The category can also include factory trawlers supporting a fleet of smaller catching vessels that are not carried on board.

Support and ancillary vessels. Large fishing fleets may be provided with support from rescue and hospital ships so that they are fully self-sufficient. Fishery protection vessels also may incorporate similar facilities, although their primary role is the enforcement of fishing rules and regulations—particularly those relating to fishery limits. These vessels are usually armed in order to ensure compliance with their requests and are often naval vessels. Training vessels and fishery research vessels are usually equipped for different fishing methods, but they have additional facilities to meet their particular functions.

Fishery protection

Terry Domico/EARTH IMAGES



Pacific salmon troller.

Freshwater fishing boats. Most of these use net and line methods rather than trawls and are therefore lighter in construction than their seagoing counterparts. They are generally small and often use unique fishing methods developed to suit local conditions. (D.Pi.)

HARBOURS AND PORT MARKETS

The purpose of a fishing harbour is to provide safety for boats, to transfer the fish rapidly to the market or consumer, and to speed the vessels' landing, maintenance, and departure. Facilities for unloading, sorting, and weighing, market halls, merchants' buildings, ice plants, refrigerated storage, and processing plants may be needed depending upon the type of fish product being handled. For vessel repair, a shipyard or slipway is necessary, with engine and electronics shops, net loft, chandlers, and victualers. Fuel may be supplied from floating or shoreside stations, and fresh water is needed for drinking, cleaning, and processing. Electric power, road and rail access, and good communications facilities are important, together with waste disposal arrangements to prevent harbour pollution.

In its simplest form, a fishing harbour may be a stretch of open beach or a landing site in a rocky coastline, from which canoes or small boats are launched and retrieved manually or with a simple winch. The catch may be sold directly to consumers from the boat or to middlemen who distribute the fish within a wider area. Facilities may include simple storage sheds for nets, outboard motors, fuel, and other supplies. A much more expensive alternative is a pier constructed beyond the surf zone, where boats are lifted out of the water by crane and stored on the deck.

Natural and man-made harbours

Larger, heavier boats with inboard engines require sheltered anchorages or moorings, since they must remain afloat. In sheltered coves or inlets a simple rock jetty or wood wharf may be sufficient, with storage, fish handling, and service facilities nearby. If a sheltered site is not available or has insufficient water depth or area, it is necessary to construct protective breakwaters to form an artificial harbour within which the necessary quays, slipways, support, and fish-handling facilities are built.

Larger harbours serving many bigger vessels may be planned and built by government bodies that operate the facility themselves or lease it to a port authority or company. Often, service facilities such as net lofts, slipways, engine and electronics dealers, and supply stores are provided by small entrepreneurs or private companies.

Fish-handling and processing facilities may include individual stalls selling fresh fish, auction halls where wholesalers purchase their supplies, or processing plants with freezing facilities producing sophisticated consumer packs from raw fish landed at the docks.

Most harbours are communal in use, serving a wide range of vessel owners, merchants, and processors. However, large companies with considerable fleets often construct private integrated facilities that include processing, maintenance, and service operations. This is especially the case in developing countries, where only small-scale landing facilities may exist. (J.C.Sa.)

Types of fishery

SALT WATER

Fishing in salt water ranges from small, traditional operations involving one person and a rowboat to huge private or government enterprises with large fleets for deep-sea and distant fisheries.

The Law of the Sea

The Law of the Sea extended from 12 to 200 miles an exclusive economic zone (EEZ) within which a coastal country has control over fisheries and their exploitation. This effectively restricts most fishing operations on the continental shelves to national vessels or to craft licensed by that country. Within the EEZ, fresh water and coastal waters are often demarcated by law, with fishing within, for example, three miles of the coast allocated only to small-scale, non-trawling fishermen and larger industrial vessels required to remain farther offshore. Small-scale fishermen are usually not restricted to the three-mile zone, and they often may be found well offshore or along the coast from their home ports as they follow the fish. For example, West African canoe fishermen traditionally migrate hundreds of miles coastwise in open canoes, frequently fishing out of sight of land. (A.v.B./J.C.Sa.)

The oceans. The oceans constitute the largest factories of living organic matter on Earth, in both magnitude and total productive biomass. Average organic production per acre is identical to that on land, although productivity varies greatly from one area to another, ranging from luxuriance to almost barren deserts. Production in any specific area varies with the seasons and is subject to large and sporadic fluctuations.

The primary production area of the oceans is the photic zone, the relatively thin surface layer, 25 fathoms (50 metres) deep, that can be penetrated by light, allowing the process of photosynthesis, the use of energy derived from sunlight in the manufacture of food, to take place. All marine life is directly or indirectly tied to the photic zone, on which both recycling and decomposition, also in other spheres of the ocean, depend. Those few microorganisms deriving their energy from sources other than light have relatively little significance in the overall productive balance of the oceans.

In the photic zone, growth rate depends on light intensity and available nutrients. Nutrients are constantly depleted by the slow sinking toward the bottom of dead plankton, the floating and mainly miniature plant and animal life, which forms the primary link in the ocean food chain. Simultaneously, fertility is constantly restored as the nutrient-rich deeper waters are brought to the surface. The ocean is ploughed by the action of winds drifting surface waters away from coastal areas, by nutrient-rich waters welling up from the depths, and during the winter season of the temperate regions by cooled surface waters becoming heavier and sinking downward, forcing nutrient-rich waters to rise.

As a rule tropical surface waters do not interchange with the mineral-supplying waters below as much as those of colder regions and are therefore less productive. However, under certain conditions in some regions of the tropics and subtropics, currents and winds induce a sustained upwelling of mineral nutrition from lower strata, producing spectacular results. Such regions include the waters around the west coasts of southern Africa and South America. Consideration of such conditions demonstrates that the production of fish-supporting plankton is not related to latitude but depends upon the presence of "new water" high in nutrient salts.

Food chains

The marine food chains, ranging from minute floating phytoplankton, sometimes called the "grass" of the sea, to the large predatory species, have many more links than terrestrial equivalents. Each transfer of food value from a lower to a higher level involves a considerable loss in the amount of recoverable organic matter, and consequently of food, so that the amount of organic matter is much greater at the plankton level than it is in fishes. The daily production of dry organic matter in kilograms per square metre beneath the surface of the English Channel is as follows: phytoplankton (plant life) 4–5; zooplankton (animal life) 1.5; pelagic fish (living near the surface) 0.0016; bottom fish 0.0010.

The plankton eaters, although they tend generally to be small in size, include the basking and whale sharks, the largest of all fishes. Typical consumers of marine plankton include such species as herring, menhaden, sardines, and pilchards. Because of this plentiful food source, these fish exist in tremendous numbers, forming the basis of important fisheries.

Demersal fishes, including such species as haddock and halibut, live primarily near the ocean floor, where they feed on various invertebrate marine animals. Most of the large fish, such as tuna, swordfish, and salmon, feed on smaller fishes.

Objects of sea fishery. *Fishes.* Small, schooling pelagic species are the most abundant fish in the near surface waters of the seas. Pilchards, capelin, herring, sardines, anchovies, menhaden, and small mackerels make up more than one-quarter of all saltwater landings. These fishes travel in immense schools several miles long and wide, containing thousands of millions of individuals. Herring feed on small marine animals and other plankton; in turn, such predators as cod, mackerel, tuna, and sharks, as well as certain kinds of whales and birds, eat freely from the enormous schools. Actual landings of each species tend to follow cycles as a result of fluctuations in the size of the resources owing to natural environmental changes and fishing pressures. Most of the catch is directed to production of fish meal and oil.

The codfishes, including cod, hake, haddock, whiting, pollock, and saithe, share with herring the leading place among edible marine fish. Alaska pollock is the most important, particularly for Russia and Japan. Atlantic cod is an important food fish in both Europe and North America.

Salmon are anadromous, migrating to ocean waters for growth and returning to fresh water for spawning. Pacific salmon return to the freshwater rivers once, to spawn and die; the Atlantic salmon make several returns. Industrial pollution, silting, and damming of rivers for hydroelectric power have seriously threatened the salmon. Only through such large-scale management measures as bypass streams and hatcheries has it been possible to save the Pacific salmon; similar measures with respect to the Atlantic salmon have been less successful.

Flatfish include a great many species, such as plaice, halibut, and sole, living largely at the bottom of the coastal shelves. The stock of each species is quite limited, however, and halibut was one of the first species for which catch quotas were established.

A major change in ocean fishery since World War II is the intense exploitation of redfish, also called ocean perch. Jack mackerel, one of the earliest fishes used for human food, continues as an important food source. Although it lives in midocean, the catch has increased.

Tuna

The true tuna fishes include albacore, bluefin, bigeye, yellowfin tuna, bonito, and skipjack. These species represent a significant marine source of human food, hunted since ancient days. Both Atlantic and Pacific stocks have been heavily fished since the end of World War II, and signs of excessive harvest have appeared. More than half the global catch is canned or frozen for the U.S. market. Spanish mackerels and swordfish belong in this group but, despite efforts toward increasing the catches, both remain minor items.

There are some 250 species of shark. Like the whale, sharks have a broad range of feeding habits. Although many are predators, some, including two of the largest fishes in the oceans, the basking shark of the northern temperate zone and the whale shark of tropical waters, are plankton feeders. Shark meat is commonly eaten in warm latitudes but elsewhere is little esteemed, except for the fins, high in protein and considered a delicacy, which are frequently used in soups.

Since World War II, many new fish species have been exploited. The clearest indication of this is the doubling of the catches of nonidentified fishes, a category that equals the volume of codfishes.

Shellfish. The term shellfish is generally applied to all invertebrate marine organisms having visible shells. They may be broadly categorized as crustaceans and mollusks.

The crustaceans include lobsters, crabs, crayfish, and both shrimp and the closely related but larger prawns. The shells consist mainly of a hard, inedible substance called chitin. Crustaceans molt frequently during growth. Blue crabs are eaten when molting and soft-shelled. Marine lobsters are eaten when about five years old and have by then molted about 25 times.

The shrimp industry

With the development of satisfactory freezing techniques in the 1940s, shrimping expanded considerably, becoming a global operation. The United States is a major consumer, importing shrimp, mainly frozen, from more than 60 countries. South Africa and Australia have developed a worldwide market for rock lobster, and Japan and Russia dominate the world market for king crab.

The major mollusks consumed as food are oysters, mussels, clams, scallops, whelks, and snails. The best-known marine snail is the abalone, encountered in many warm waters. This group also includes the octopus, squid, and cuttlefish, popular seafoods in Mediterranean countries and the Far East.

Sea cucumbers (holothurians), or sea slugs, are usually marketed under the name of trepang or bêche-de-mer. Rich in protein, they are eaten in China, Southeast Asia, Australia, and Italy.

Mammals. Ocean mammals include such cetaceans as whales, porpoises, and dolphins, as well as seals and walruses. Whales are a source of meat, fats, and oils, hormones such as insulin, and chemicals. They exist at all levels of ocean food chains. The blue whale mainly devours small reddish shrimp called krill, while the formidable killer whale feeds on salmon, seals, and sharks. The number of species, although still large, has declined considerably. See the section *Whaling* below.

The hunting of porpoises and dolphins preceded whaling in history. Dolphins were eaten in ancient times around the Mediterranean, and Xenophon and his Greek army found sizable stores of salted dolphin meat in earthenware vessels on the Black Sea coast. Their use as food there continued until banned by the Soviet Union in the interest of preserving the animals for biologic research. Many tropical islanders still hunt dolphin on a large scale. Freshwater dolphins are caught in many of the world's great rivers, including the Ganges, Indus, and Brahmaputra, the Amazon, and the Río de la Plata. The dolphins of Chinese rivers have been eradicated, but a number survive in the lake regions of the upper Yangtze.

Seaweeds and plankton. Marine plants may be divided into two groups: grasses and algae. There is only one subaquatic grass of any significance, namely eelgrass. Algae that grow in a fixed location, generally called seaweeds, may be categorized according to colour, into green, brown, red, or blue-green. Brown algae, sometimes called kelp, may grow to exceptional sizes; some specimens attain a length of 50 metres or more.

Seaweeds are heavily exploited in many parts of the globe for human as well as animal food. Several species are extensively cultivated on the coastal shelves of China, Japan, the Philippines, and elsewhere. Brown species in particular are harvested in Japan and made into a number of food products. Several are used as material for various thickening agents.

Cultivated red seaweeds belong to the genus *Porphyra.* Their sun-dried, blackish fronds are shaped into sheets and used in the Orient as a wrapping for rice. Harvested along the coasts of Ireland and Scotland, red seaweed is made into a powder and used as the main ingredient of a kind of bread called laver. Seaweeds contribute to the diet accessory nutrients such as vitamins B_6 and B_{12}.

Phyto-
plankton
as food

Phytoplankton does not offer man a suitable food and can hardly be used even as feed for animals. Many species are toxic; the rest are scarcely digestible. In addition, most plankton finish their life cycle within a few days or weeks and are usually devoured by predators. Consequently, the amount of plankton in the water at any given moment is small, even though total plankton production over a year may, in a particular water, well exceed that of fish. Plankton harvesting is therefore very difficult, because of the volume of water that must be sieved, but several attempts to develop a feasible harvesting device have been undertaken. The Japanese, Burmese, and East Indians have managed to develop profitable fisheries for certain tiny shrimp that feed on plankton. The shrimp are dried or fermented into pastes. Elsewhere similar plankton-fed shrimp are sun dried and sold as a snack.

Unicellular green algae, such as *Chlorella* and *Scenedesmus,* have been artificially cultivated, yielding 75 tons per hectare (30 tons per acre) per year, compared with the standard wheat yield of 2.5 to 3.7 tons. However, the process is costly, since algae, in addition to harvesting, require decolorization and special processing to remove or break down the cell walls through drying and enzyme action in order to become digestible. It is far more efficient to use such plankton directly as fish feed in cultivation ponds or in the raising of cattle and poultry. Blue-green algae easily create waterblooms, slimy accumulations that may be dried in the sun and molded into small loaves with a nutlike flavour and high in protein. This food is extracted from Lake Chad in tropical Africa, and the Aztecs made a similar product. In China a scum called *lan,* collected from ponds and freshwater lakes, provides sustenance for large numbers of people. A related scum, *keklap,* found in Java, is used chiefly as fish feed. Another species is made into dried sheets in Japan and prepared for food by heating in water. Successful cultivation of some blue-green species has been carried through on a semicommercial scale. (G.A.B./J.C.Sa.)

Spotting
pelagic fish

Fish finding. Traditionally, sea fishermen have known the time and place to find their catch, but the history of fishing has demonstrated more than once that even old and rich fishing places can become exhausted quite suddenly. This is especially true with pelagic fish like herring, pilchards, or sardines. The herring yields of the Schonen fishery and later on of the Bohuslaine fishery (1744–1809) in the Baltic Sea fell so severely that the very existence of the Hanseatic League was compromised. This sudden change did not result from overfishing but was caused instead by natural fluctuations in the development of stocks. In modern times, sardine fishing collapsed off the California coast in 1952, followed by the Peruvian anchovy fishery 20 years later. Similar disasters have occurred in other parts of the world not only because of overfishing but also for natural reasons. When this happens new fishing places must be found. It is difficult to explain how good fishing places in great depths were found in ancient times, but fish in shallow waters, fjords, or small bays can easily be seen. On the high seas, fish can be located when they surface temporarily, and fish searching by direct observation from a vessel is important even today. Airplanes and helicopters are commonly used in purse seine fishing. An experienced air-spotting pilot can detect fish under the surface and identify species by observing the shoal's form or colour or behaviour and sometimes by the presence of accompanying birds. During the night, fishes can be located through the phenomenon of bioluminescence; *i.e.,* when their passage through the water causes tiny marine organisms to luminesce. Accompanying birds have played an important part in fish searching for centuries, because a concentration of birds can be seen from a distance. Very often the birds are not attracted by the fish sought but by smaller fishes and squid, which may have taken refuge from large species by swimming to the surface. Other animals may also indicate fish concentrations by their presence. Porpoises, for example, are known companions of tuna, and tuna purse seiners often set their nets where porpoises have been seen. To find fish in deeper waters by other means was difficult if not impossible in the past. Herring fishermen used signal lines to find their prey in deep waters. These were long wires dropped from a boat; the fisherman holding the line in his hand could feel the vibration caused by the fish touching the line, which was named the herring's telephone. Other fish were also found by signal lines, often tied with fishing gear. In modern industrial fisheries, experiments have been made with direct listening for fish, but this method has been found impractical. Sea fishermen have also learned to judge where fish can be expected by observing environmental conditions. The colour of the water and the presence of current or of a borderline between different water bodies are some common fish indicators. One of the most important physical properties for fish finding is the temperature of the water. The use of thermometers was one of the first practices fishermen learned from oceanographers, not only for fish finding but also for forecasting availability of the desired species. Aerial and satellite surveys of these properties are becoming of increasing importance.

Echo
sounding
and sonar

The first experiments using electrically generated sound pulses and their echoes to locate fish were undertaken in Britain during the 1930s, and by the 1950s fish-finding echo sounders had become an essential aid to catching. As these units worked vertically, they only showed fish immediately below the vessel, so that a logical development was the application of sonar in order to search horizontally around the boat. For many years the machines provided only a "black on white" paper display of the resulting echoes, and interpretation of the displays was dependent upon the skill of the skipper. Gradually, improvements were made in the quality and quantity of information displayed, enabling monochrome signals to be displayed on a television-type screen. A big breakthrough came with the advent of microprocessor technology, which made it possible for fish-finding sonars and sounders to rapidly analyze the signals that their high performance transducers picked up from the sea. Information regarding size, abundance, and movement of the fish is now displayed in many colours, provides the skipper with a wide range of scales, and enables him to focus on and expand information at a particular depth or location.

Other instruments have become vital to fishing operations, especially radio- and satellite-transmitted position-fixing equipment such as Decca Navigator, Loran, and Satnav. These enable a skipper to return to the precise position where fish are spotted or to a particular location such as a coral reef or where gear has been set. Microprocessor technology allows information from various instruments such as sonar, radar, Satnav, and Loran to be fed into a single television screen that provides the skipper with information processed to suit his needs. The vessel's movements, shown on the screen, can be integrated with navigation and fishing charts fed into the display from computer memory banks. By linking these instruments to the control of winches, engines, and rudder, fully computerized fishing operations are possible.

Sea-fishing methods. *Pole-and-line fishing.* Line fishing at sea is very popular, not only in traditional fisheries with small boats employing a limited number of hooks but also in industrial operations with large vessels or fleets using thousands of hooks.

Pole-and-line methods are used in tropical Pacific and Atlantic waters to catch young bluefin and yellowfin tuna, and smaller tuna species—such as albacore, skipjack, bonito, and little tunny. The pole, generally bamboo, ranges in length from two to 10 metres, with a line of roughly the same length. Hooks of various sizes are barbless to facilitate baiting and removing the captured fish. To hold onto the pole a "rod rest" is generally used, which is made of canvas, leather, or old rubber tires. Depending on the size of the vessel, the crew may number 30 or more. A large crew is needed, since fishing time may be limited and the maximum possible number of rods must be worked. If larger and heavier fishes are sought, two, three, or even four poles may be linked to a single hook. In this case the fishermen must cooperate closely. Also used successfully are deck- and rail-mounted automated fishing poles operated hydraulically and electrically. The fibreglass rods are mechanically moved up and down, swinging the hooked fish onto the deck and removing the hook before swinging it, unbaited, back overboard.

Automated tuna fishing

By courtesy of the Federation of Japan Tuna Fisheries Co-operative Associations

Automated pole-and-line gear on a Japanese tuna vessel. One man can work as many as eight rods.

The tuna is attracted and kept near the vessel by chumming, throwing live bait overboard. The bait is kept alive on board in special tanks in which seawater circulates constantly. Bait can be an expensive problem for tuna fishermen; to catch one ton of tuna, roughly 100 kilograms of live bait fish are needed. Sometimes the hooks are baited, sometimes artificial lures are used with hooks hidden in feathers. When the tuna is "hot" (very eager to take the bait), a naked hook is sufficient. Water spraying helps to attract the tuna; it also serves to camouflage the shadows of boat and crew.

Pole-and-line fishing for tuna is done in daytime from slow-moving vessels. Since considerable space is needed for the angling crew to stand side by side on the lee side of the vessel, Japanese vessels for pole-and-line fishing have a long extended bow. To simplify hauling in the catch these boats also have a low freeboard (*i.e.,* their sides ride low above the water). American tuna vessels hang special crew racks outside the ship over the water.

Drifting longlines. Used for tuna—especially in Japan, Taiwan, and Korea and to a limited extent in South Africa, Cuba, and Oceania—drifting longlines are particularly successful in the tropical Atlantic for big fish in depths from 60 to 250 metres. More than half the fish caught in this manner are yellowfin tuna, one-third are albacores, and the remainder bigeye and bluefin tuna. Sharks, marlins, swordfish, and sailfish, also caught with drifting longlines, are sometimes included in the tuna statistics. Sharks can cause serious losses by attacking hooked tuna. Originally longlining for tuna was a Japanese inshore fishery. At the end of the 19th century, the Japanese were fishing 50 to 65 kilometres off their coasts. This fishery was extended when sailing boats were replaced by motorized craft, and by 1926 the Japanese began longlining for tuna off Taiwan, by 1929 in the Indian Ocean, by 1930 in the South Pacific, by 1938 in the eastern Pacific, by 1952 off the southeastern coast of Australia, and since 1955 in the Atlantic. A longlining crew must be willing to do a hard, though lucrative, job and remain far from home for long periods. The gear is a line composed of 400 to 450 sections, each section with a length of 150 to 400 metres stored in a basket. The total line can have a stretched length of up to 180 kilometres. Each section is composed of subsections of different length. The branch lines with the hooks are composed of three sections that vary in number and length. From one to 12 (generally five) branch lines with hooks form one section; 2,000 hooks are considered the greatest number that can be operated in one set by a vessel. With decreasing catches, attempts have been made to increase the number of hooks; Korean fishermen are said to operate as many as 3,000. The shooting of the line from the stern of the vessel begins early in the morning before sunrise, when the vessel is moving at a speed of about five knots (five nautical miles per hour) or more. During shooting the lines have been tied together and the hooks are baited with frozen Japanese sauries. Each section is tied with a float line and a buoy. Depth of the gear can be regulated by the length of the float lines and the distance of the floats. Ten to 14 men require four hours to perform the task. Hauling from the forepart of the vessel begins in the early afternoon with the help of a line hauler. Depending on the quantity of the catch, hauling can take more than 10 hours with a crew of eight to 10. With preparing and sorting the catch, the usual working day of a crew member totals some 18 hours. Because of this and the fact that vessels stay at sea more than 200 days per year, the Japanese and Taiwanese have experienced difficulty in procuring crews; this problem has led to the development of new technology to simplify the work and reduce manpower. One such improvement is the reel system, made especially for larger vessels. The total line is set, hauled, and stored on a drum, and the floats and branch lines are stowed on separate reels and clipped on or removed as the main line is set or hauled. Research is being done on a coupling apparatus to do this automatically. Another invention is a line-winder system practicable for small vessels. In this a single line is used, hauled and coiled by a line winder in special tanks in the aft part of the vessel.

Labour-saving technology

Bottom longlines. For centuries, line fishing for demersal fishes was carried on in coastal waters and far at sea in the dory fishery famous today. A sailing mother ship carried the dories from Portugal, France, Canada, and the United States to the Grand Banks for cod. The one-man dory operated near the carrier setting longlines and sometimes fishing with handlines. In the evening the catch was carried back to the mother ship where each man prepared his catch for salting. Some large-scale modern enterprises also fish with bottom longlines, catching many species of the cod family, including cod, haddock, coral fish, hake, and pollock, as well as rays, and many flatfish, such as halibut. There are also longline fisheries for groupers, hairtails, croakers, and sea breams. Bottom lines are not as long as the more easily controlled drift lines. The hooks do not always lie on the bottom but may hang above it to protect the bait against unwanted bottom predators, such as starfish, snails, or crabs. Typically, bottom lines are used for halibut in the northern Pacific. A relatively heavy main line is divided into sections of approximately 90 metres. The branch lines, each about 1.5 metres long, are tied at intervals of four to 5.5 metres. Modern synthetics, with their greater strength and lighter weight, have replaced natural fibres for main lines. Fishing depth usually ranges between 80 and 270 metres, depending on the grounds and season. The setline is anchored on both ends, marked by a floating keg and a lighted flag buoy at night.

An automated longline system developed in Norway baits hooks when setting, then cleans and stows them on magazine racks when hauling. This, and a number of similar systems, has enabled more hooks to be set by smaller crews and has thereby revolutionized the bottom longline fisheries of Europe and North America.

Traps. There are only a few areas in the world where water or weather conditions prohibit the use of traps. A single small vessel can operate hundreds of traps, though lack of storage space may cause difficulties. Thus collapsible traps of netting on a wire framework are often preferred not only for fish but also for crustaceans. Many plastic traps are made, especially for lobster. Some can be dismantled for easy transportation. Water snails, such as whelk in England and other species in Korea, are also trapped, as are cuttlefish and octopuses. As in fresh water, fyke nets can be set in long rows or in connected systems.

Setting traps

Commercial sea fisheries set long rows of pots or framework traps by the longline system; *i.e.,* single pots are tied with a branch line to a main line. Hauling is accomplished with small hand-operated or motor-driven winches. More important for catching fish in commercial sea fisheries are the big wooden corrals, or weirs, and the large pound nets. The oldest type may be the Italian tonnara, used in the Mediterranean for tuna from the Bosporus to the Atlantic. Very large pound nets are also used by the Japanese on the Pacific coast, by the Danes and their neighbours off the eastern coasts of the Baltic, and for salmon fishing off the Pacific and Atlantic coasts of North America. The difficulty in setting large traps lies in placing them on the bottom. If the water is not deep and the bottom is not hard, the weirs can be held by sticks or piles. Where the water is deeper and the ground is hard or rocky, the weirs must be anchored.

Dragged gear. Dredges and trawls are of great importance in commercial sea fisheries. Dredges are generally used in shallow water by small vessels, although a deep-sea dredge is operated by research vessels at depths of up to 1,000 metres. The simplest dredges in sea fishery are hand operated. Fitted with a stick up to five metres long, they resemble rakes combined with a bag for collecting the catch—usually mollusks or crustaceans. Heavier dredges with a triangular or quadrangular iron frame may be towed along the seafloor by small vessels or pulled some distance from the shore or from an anchored vessel and then towed back with a winch. For digging out mollusks, some dredges have iron teeth on the lower edge of the frame. They may also have a pressure plate on the upper part and chains on the lower part, depending on the catch sought. The bag of the dredge is made of wire rings that have good resistance to friction and of hard fibre netting. Usually more than one dredge is operated by a vessel, and they are towed with the help of outriggers. The great disadvantage of dredging is that much of the catch is damaged, wasting effort and needlessly killing fish.

Trawling

Trawling in sea fishery can be done by small vessels or even rowboats (as in the estuary of the Tagus River near Lisbon). More important, however, are fleets of highly mechanized trawlers whose gross registered tonnage may reach 5,000 and whose horsepower approaches 6,000. The trawl is a towed net bag with a wide opening at the mouth and an end closed by a special knot. The mesh size of the opening can be large—600 millimetres (two feet) from knot to knot—to diminish water resistance during towing. The closed end (called the cod end) can have meshes of six millimetres, depending upon the species of fish or shrimp sought. The trawl is designed in a smooth funnellike shape to guide the fish into the cod end. To keep the mouth of the trawl open, a large horizontal beam may be used. The beam can measure up to 12 metres in length and is based on two guides that glide over the bottom. The Dutch catch flatfish with beam trawls that have heavy chains, called tickler chains, dragging on the seafloor in front of the net opening between the two gliders to frighten the fish from the bottom into the trawl. Additional stimulus is often provided by electrifying the tickler chains.

Though beam trawls were the original gear of deep-sea steam trawlers, today they are used by smaller vessels only. Beam trawls are usually towed in pairs, one on each side of the vessel. Such an arrangement can considerably decrease the stability of the vessel and is dangerous except in craft specially designed for the purpose. Another method involves two vessels stretching the horizontal opening of the trawl between them. Two vessels have more power to tow a bigger trawl at greater speed, but the skippers of the two vessels must cooperate very closely. The most important method for spreading a trawl opening employs

(Left) Dag Pike, (right) David Falconer/After-Image

(Left) Dutch outrigger trawler hauling a heavy beam trawl. (Right) Cod being hauled aboard a Russian stern trawler in the Bering Sea.

two trawl doors, or otter boards, rectangular or oval plates that are attached to each side of the net and caused to flare apart by the pressure of the water.

Mid-water trawling involves dragging the trawl with one or two vessels in the area between the ocean bottom and its surface to catch pelagic fish. The trawl is set at the depth where fish have been observed by varying the length of the towing warps and the speed of the towing vessel. With longer warps and lower speed, the trawl sinks; it rises with shorter warps and higher speed. The depth of the trawl is monitored by a special transducer called a netsonde, which is mounted on the trawl and transmits echograms showing the position of the net in relation to the bottom and to the school of fish.

The netsonde

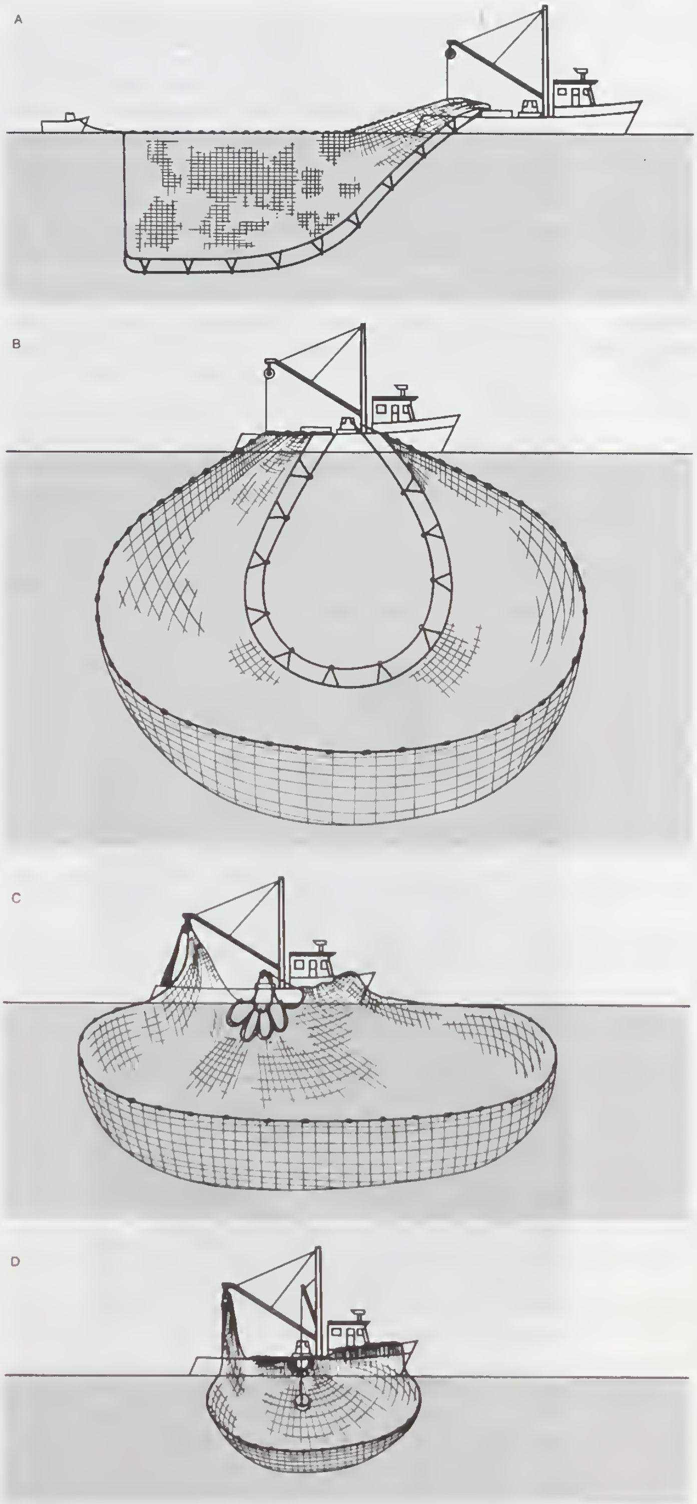
A special type of mid-water trawl is the semipelagic trawl, originally invented in Iceland and now operated primarily by French fishermen. In this technique the otter boards remain in touch with the bottom but the trawl floats at some distance above it. Semipelagic trawls were constructed because fish often are concentrated at a short distance from the bottom outside the range of the usual bottom trawl, which has a low, wide opening. To overcome this difficulty, a higher opening of the trawl is needed. Though the opening of a bottom trawl can be stretched vertically by various means, such stretching decreases the horizontal width of opening. Some modern bottom trawls are constructed with a high vertical and horizontal opening, and many consider them the best available gear for bottom trawling.

Seine nets. Seine nets are often employed in beach seining, where fish shoals are near beaches. Large beach-seining operations for sardinelike fishes and other species are carried on in the Indian Ocean. The importance of this method has decreased as pollution has cut the available stocks of fish in this region and as manpower costs have risen: not all fishing methods lend themselves to mechanization. More successful are anchor seines, better known (because of their origin in Denmark in 1849) as Danish seines. The gear consists of a net similar to a trawl but with a large bag and long wings connected to long towing ropes. One of the ropes (up to 1,000 metres long) is tied to an anchored buoy. The other rope is tied to the vessel, which steams in a wide circle, laying the ropes and returning to the buoy. The ropes act to keep the net open and herd the fish toward the bag. The vessel then hauls both ropes together until the net bag is taken on board. This method is used in northern Europe for flatfish and cod and in Japan has become the most important method of inshore fishery for bottom fish, after two-boat trawling.

Purse seines and lamparas. The most important sea-fishing gear is the surrounding net, represented by the older lampara nets and the more modern purse seines. Both are typical gear for pelagic fish schooling in large and dense shoals. When these nets are used, a shoal of fish is first surrounded with a curtain or wall of netting that is buoyed at the surface and weighted at the bottom. The lampara net has a large central bunt, or bagging portion, and short wings. The buoyed float line is longer than the weighted lead line, so that, as the lines are hauled, the wings of the net come together at the bottom first, trapping the fish. As the net is brought in, the school of fish is worked into the bunt and captured. With the purse seine, once the school is surrounded, the bottom of the net is closed by drawing a line through rings attached to the lead line. This pulls the net shut at the bottom like a purse, and when the net is hauled in, the concentrated fish are removed by a brail (dip net) or are pumped aboard the fishing vessel.

Hauling a purse seine

Surrounding nets are used for tuna, herring, sardines and related species, salmon, mackerels, and even cod (when they come to spawn in the pelagic zone). For these nets to be successful, the fish must be in large and dense shoals; light and bait are sometimes used as lures to produce such shoaling.



Setting and hauling a purse seine.
(A) Attaching one end of the seine net to the skiff, the seiner steams around a school of fish while laying the net. (B) Seiner hauls in purse lines, drawing shut (pursing) the bottom of the net. (C) Seiner hauls float lines, drawing in the web of net. (D) Fish are concentrated alongside, where they can be pumped or brailed aboard.

After J C Sainsbury, *Commercial Fishing Methods, An Introduction to Vessels and Gear*, (1971); Fishing News Books Ltd

Lift nets. Fish can also be caught, in limited quantities, by lift nets: stationary types operated along the shoreline, movable ones from rafts and boats, and large blanket nets held on each corner by a small boat. The Soviets operate a large commercial lift-net fishery on the Caspian Sea to catch sardinelike fish attracted by light. Each vessel operates two conical nets, setting one while the other is being lifted. Another effective lift net is the large, boxlike *basnig* of the Philippines, operated with a luring light during the night beneath a single outrigged vessel; sardines, mackerels, hairtails, squid, and other pelagic prey are caught. The Japanese have a special kind of lift net for sauries; the fish, attracted by light, swim over the netting lowered into the water and are caught when the netting is hauled.

Gill nets and drift nets. Quite important in commercial sea fisheries, gill nets are sometimes operated in large sets thousands of metres long. These generally drift with the vessel or are set as anchored nets in long rows at or near the bottom of the sea. Gill nets are used for many pelagic fishes, such as herring, pilchards, sardines and related species, mackerels, croakers, salmon, and tuna. They also are used for many bottom fishes—cod, Alaska pollock, and others. For cod, Icelandic fishermen set up to 90 nets, each about 50 metres in length, in depths up to 180 metres.

Historic importance of drift nets

Drift nets are widely used to catch pelagic sea fishes. In northern Europe, before the introduction of trawling, drift nets were the most important method of deep-sea fishery. In the old herring fishery of northwestern Europe, drifters commonly set more than 100 nets, each about 30 metres in length. Thus a fleet of drift nets might measure three or even four kilometres. The nets are set in the late afternoon to catch the herring as they ascend in the evening from ocean bottom to higher water levels. During the night the vessel drifts with the nets like a buoy. Hauling, done by hand or with mechanical aids, begins at midnight and, when big catches are taken, can continue until late morning. The fish are shaken out of the meshes by hand or with shaking machines.

© Terry Domico/EARTH IMAGES

Hauling in a gill net loaded with salmon.

Entangling nets. Similarly operated are entangling nets, single or double walled, and three-walled trammel nets. These are used in sea fisheries for hake, shark, rays, salmon, sturgeons, halibut, plaice, shrimps, prawns, lobster, spiny lobster, king crabs, and turtles. Single-walled nets are used in the southern part of the Caspian Sea and in the Black Sea to catch sturgeons by entangling. Iranian fishermen set about 150 sturgeon nets in one row perpendicular to the shoreline. Setting requires much labour; between each two nets a line is tied, which is connected to a short wooden peg driven into the bottom. The Turkish Black Sea fishermen sometimes set sturgeon nets in another form. Two nets always form an angle open to the sea. The nets are held by sticks rammed into the bottom. Sturgeon nets are checked once or even twice each day, depending on weather. For this purpose an Iranian fisherman lies on the bow of his sailboat, towing the vessel along the float line of the net. The sturgeons are taken from the water by hand or with a gaff.

The most important sea fishery for crustaceans is the king crab fishery in the northern Pacific. For the Japanese, who use entangling nets, this is a very important distant fishery ranking with tuna and salmon fishing. Originally carried on close to shore, king crab fishing was extended in the northern Pacific after its beginnings in the 1870s. The old land stations for processing were replaced by floating factories that accompanied the fishing vessels. The entangling nets are set on the bottom, sometimes 200 nets with a total length of 10 kilometres in one row. Larger catching vessels set 1,200 to 1,300 nets a day, usually in parallel rows about 500 metres apart. Nets stay in the water from five to seven days and are hauled by small open vessels with motor-driven reels, which can take from 2,500 to 3,000 nets per day out of the water. When hauling, the floats and sinkers are untied and the entangled king crabs are taken from the netting. The catch and nets are then transported to the mother ship, where the catch is processed and the nets cleaned, an operation that may require 30 minutes per net. Large racks for drying and cleaning the entangling nets are characteristic of this type of vessel. A single fishing unit may own a permanent set of 15,000 to 30,000 nets.

Squid fishing

Harvesting machines. A relatively new type of fishing gear is the harvesting machine combined with a pump, used in the northern part of the Caspian Sea for sardinelike fish and for squid off the California coast. In both cases the prey is attracted by light. Squid fishing can be done near the surface, but in the Caspian the fish are sucked on board with pumps from depths as great as 110 metres. In pumping, the suction nozzle is moved up and down with attracting lamps. Once on board the fish or squid are strained from the water. The difficulty in fish pumping is to avoid damage to the catch. Only small objects can be pumped without injury.

Another type of harvesting machine is the hydraulic dredge, with pumps and conveyors. These dredges wash out deeply buried mussels with jets of water under high pressure. The Americans operate such hydraulic dredges to harvest soft clams, and the British use similar machines for cockles. Harvesting machines also are used to cut kelp off California. Giant kelp is harvested by cutting to a maximum depth of 1.2 metres below the surface of the water and is transferred by conveyor belt into the open hold of the vessel. (A.v.B./J.C.Sa.)

FRESH WATER

Freshwater fishing is carried out in lakes and rivers or streams and to a growing extent in natural and artificial ponds. In some tropical areas, swamps with shallow water, sometimes overgrown with vegetation, are important inland fisheries. Before efficient transportation and distribution of ocean fish was organized, fresh waters were the only resource available for fish and other aquatic products for the inland population. Their importance decreased with the growing bulk fisheries of the seas. Freshwater fish now compose only about 5 percent of the total catch of water products of the world.

Major freshwater species

General characteristics. Widely different freshwater species—feeding on bacteria or detritus, plants or plankton, or living as predators—are used for human consumption. Well-known species include trout and whitefish, carp and other cyprinids, catfish, murrals, and tilapias. The desirability of some anadromous fishes—those, such as salmon and sturgeon, that spawn in fresh water but live in the sea—and catadromous fishes—those, most notably the eel, that spawn in the sea but live in fresh water—has led to specialized fisheries in inland waters.

The kind and quantity of fish found in lakes and rivers vary greatly with the physical and chemical condition of the water. Limnologists, scientists who study conditions in fresh water, classify fresh waters by the quantity of oxygen and essential nutrient salts (nitrates, phosphates, and potash) they contain. Fishermen classify waters by the principal fish to be caught therein. Rivers, for example, are divided into different zones beginning with the source, which is often good trout water, and ending in the estuary, where many coastal varieties of ocean fish can be caught. In like manner, fishermen classify lakes by expected catch (*e.g.*, eels, tilapias, or crayfish).

The great variations in the productivity of inland waters are explained by differences in their physical and chemical properties. Though some rivers may produce as much as 200 kilograms per hectare (180 pounds per acre) each year and some lakes may yield 160 kilograms per hectare, the world average is about eight kilograms per hectare.

Pollution produced by chemical preparations applied for agricultural purposes has created serious problems for the world's freshwater fisheries; fish cultivation is increasingly restricted to man-made waters. Traditional freshwater fisheries still supply basic protein to China, Southeast Asia, and tropical Africa but have been seriously affected in the United Kingdom, continental Europe, Japan, Central Asia, and the United States.

Because of pollution, freshwater fishing in natural waters has declined in industrial countries, but pollution is not totally to blame. The rapid rise in angling as a leisure pastime has created competition for the available waters and the fish in them. Because angling interests can afford higher prices for the rights to available waters, angling is now virtually the only fishing for wild fish that takes place in natural waters in industrialized countries. Some fish species that are considered delicacies and attract high prices are exempt from this trend. Fishing for salmon, eels, and crayfish is still very active on a commercial basis. With these fisheries there are many traditional rights to fishing certain waters.

In nonindustrialized countries freshwater fishing has increased considerably, mainly under the influence of aid programs. Some of these programs have tried to introduce new and more efficient fishing methods, but the main improvement has been in mechanization of the fishing boats used and in improved methods of preserving and distributing the catch. On some of the larger inland lakes, freshwater fishing is still the primary occupation in the villages along the shore.

Fish farming for freshwater species is being introduced in developing countries to produce a valuable source of protein. Where natural waters are fished in developing countries, fish management techniques are being used to improve the catch and to prevent overfishing.

Methods. Many techniques are employed to catch fish in inland waters, some appropriate to lakes alone, some to rivers only, and some to both. Of the many methods employed worldwide, only a few are economical for large-scale operation. Commercial line fishing, which uses many hooked and baited branch lines tied to a single main line, is widely practiced. A simpler technique is handlining, in which single lines with baited hooks are tied to small sticks or trees along the shore or to special devices set along the side of a hole in ice. Handlining is used for deep fishing or for catching in rocky areas. Drifting lines with one or more hooks can also be used on lakes, though seldom in rivers. Lines may also be trolled (trailed) behind a moving boat. On some rivers in tropical and temperate areas, fish are caught by fouling with sharp-pointed hooks. The main difficulty in line fishing is to keep the lines clear and to obtain baits in needed quantity.

Passive and stationary fishing gear

Passive and stationary fishing gear is so important in many lakes and rivers that some fishermen specialize entirely in trapping. Since deep and rocky shores, however, do not favour the use of traps, these devices cannot be used in all areas. Fish seeking shelter may be caught in simple brushwood devices when the brushwood is lifted quickly. More important are traps, such as wooden baskets made of wickerwork or of split bamboo, with retarding devices such as funnels or valves at the entrance. Wooden baskets, generally used in rivers with strong currents, can be set according to the longline system in which the baskets are tied with branch lines on a main line lying across the bottom of the river. Such baskets are usually baited, with the bait sometimes held in small bags or boxes. Today, in river fisheries as in coastal sea fisheries, traps, especially those used for eels, are made of plastic.

A more modern type of trap is the bag-shaped fyke net, held open by hoops; linked together in long chains, these are used to catch eels in rivers. When equipped with wings and leaders, fyke nets are employed in lakes where there are sheltered places with abundant plant life. Hundreds of such nets can be combined into systems where it is not economical to build large traps.

Another fishing method important in freshwater fisheries employs small scoop nets or large net bags (stownets). Such gear is known on many European and Asian rivers. The net bag is fixed to the river bottom to catch migrating or drifting fish. Some human control may be necessary; sometimes a watchman lives on a vessel or raft next to the stownet or on a special platform. Though stownets are especially popular in European rivers for eel fishing, their importance is lessening owing to increased boat traffic and to pollution. Moreover, the gear can be too large to be moved easily. In Indonesia stownets up to 100 metres long and up to 40 metres across the mouth are used. Small scoop nets can be operated by hand and pushed or towed over the bottoms in shallow waters. Sometimes this is done by fishnetting parties, in which all the men of a village form a line across the river, with a scoop net in each hand. Sometimes the fisherman stands on a platform built on the side of the stream and simply scoops up fish as they pass; this is done by some African fishermen in Malaŵi and was done by American Indians on the Columbia River in Oregon.

Seining

One of the most common fishing methods in freshwater lakes and rivers is seining, which is done in temperate zones especially in autumn and winter when fish are concentrated in deeper parts of the lakes. Because part of the seine must be dragged over the lake bottom while it surrounds the fish, seining is practicable only where the lake bottom is smooth and where favourable areas (*i.e.*, with fish concentrated near the bottom) are known. In some lakes such areas have been known for a long time and may be named and marked on fishing maps. Seine nets in lake fisheries can be very large, with wings of 1,000 metres each. Since traditional seining required considerable labour, mechanization became desirable. A modern mechanized seine-net fishery requires only a small labour force. In northern countries seine nets are used under ice. For this purpose a number of ice holes are needed for guiding the towing warps with the net on the underside of the ice sheet. Here also manpower is saved by motorized towing and coiling lines and by drilling the holes in the ice with power drills. Some success has been achieved in increasing the efficiency of seine nets by electric light. Fish trying to evade the net can be caught by stunning, or eels lying in the mud during the cold season (generally a time when eel fishing is poor) can be attracted out of the mud by an electrical current. The disadvantage of all seine nets is that they are not selective; many undersized fish that should be preserved cannot escape.

On larger lakes, sea-fishing methods such as trawling and purse seining are used. Two or more fishing boats are usu-

Cary Wolinsky/Stock Boston

Catching sturgeon on the Volga River, Russia.

ally required to set a purse seine net, which can then be hauled in manually by people on the shore. Lift nets are often used in fresh water, not only to catch bait fish for line fishing but also to catch crayfish or other freshwater crabs. There are small hand-operated lift nets tightened by frames and larger ones lifted from a gallows or with one or two vessels. Unframed blanket nets are used in rivers in Italy, each corner held by a gallows placed on the banks.

Cover pots and cast nets

Cover pots and cast nets also have some importance in commercial freshwater fisheries. Cover pots are especially used in rice fields or shallow waters with rich vegetation. Cast nets are used more in clear waters in lakes and in rivers; considerable skill is required to cast these. In Russia shooting mechanisms are employed to cast larger nets. Much more important in freshwater fisheries, however, are gill nets. The mesh size of the net can be used to regulate the size of the fish caught; thus smaller, undesired species escape. Lake fishermen use mostly stationary gill nets, anchored near the bottom or floating. River fishermen use gill nets that drift with the current, with one side tied to the boat and the other to a drifting buoy. With entangling two-walled nets and trammel nets, yield can be increased by frightening the fish into the netting; this is accomplished by beating the water or throwing stones. (A.v.B./D.Pi.)

AQUACULTURE

Aquaculture is the propagation and husbandry of aquatic plants and animals for commercial, recreational, and scientific purposes. This includes production for supplying other aquaculture operations, for food and industrial products, for stocking sport fisheries, for producing aquatic bait animals, for fee fishing, for ornamental purposes, and for use by the pharmaceutical and chemical industries. These activities can occur both in natural waters and in artificial aquatic impoundments.

Aquaculture has been in existence since at least 500 BC. However, only in recent times has it assumed commercial importance, with world production more than doubling between 1970 and 1975. The rapid expansion of aquaculture has been to a large extent in the production of relatively high-priced species frequently consumed as a fresh product. Examples are shrimp, crayfish, prawns, trout, salmon, and oysters. However, also increasing is the production of catfish, carp, and tilapias, which are reared in extensive, low-energy systems. For example, catfish farming in the United States has more than quintupled its production since it began to grow in the 1960s.

The growth of world aquaculture has been stimulated by a number of factors, including population increases, dietary shifts, and advances in aquaculture technology. Limited ocean resources have also helped to create a growing role for aquaculture in helping to meet increasing demands for fish and shellfish.

First artificial fertilization

Farming and rearing in hatcheries. Fish farming as originally practiced involved capturing immature specimens and then raising them under optimal conditions in which they were well fed and protected from predators and competitors for light and space. It was not until 1733, however, that a German farmer successfully raised fish from eggs that he had artificially obtained and fertilized. Male and female trout were collected when ready for spawning. Eggs and sperm were pressed from their bodies and mixed together under favourable conditions. After the eggs hatched, the fish fry were taken to tanks or ponds for further cultivation. Methods have also been developed for artificial breeding of saltwater fish, and it now appears possible not only to rear sea animals but also to have the complete life cycle under hatchery control.

Carp. Carp raising, practiced worldwide, is a good example of advanced techniques. For the whole life cycle at least three different types of ponds are used in Europe. Special shallow and warm ponds with rich vegetation provide a good environment for spawning, a process that today is often aided by hormone injections. After spawning, the parent fish are separated from the eggs and taken to a second pond. The fry, which hatch after a few days, are transported to shallow, plankton-rich nursing ponds, where they remain until the fall of the year or the next spring. In tropical areas, such as India, carp spawned from wild fish can be collected by experts in natural waters. To collect eggs or fry from wild fish is disadvantageous, however, because the breeder cannot influence the breeding stocks in a desired direction. In Asia, the fry of common or golden carp are thus generally bred under culture conditions in hatcheries. Bigger ponds are needed for rearing the fish in the second year of life. There are large carp ponds in certain areas of central Europe, while in Asia common carp are often cultivated in rice fields, a practice called wetland cultivation. This method is increasingly jeopardized by sprays used to control pests and diseases and by toxic agents resulting from industrial development. For feeding carp in ponds, soybean meal, rice bran, and similar agricultural products are used. Concentrated food in the form of pellets has also been successfully introduced. During the winter season in the temperate zone, the carp are kept in deeper ponds with a dependable flow of water to protect them against freezing. In central Europe, carp are ready for the market after the third summer. In southern Europe, Hungary, and parts of the Balkan Peninsula, carp may be sold after the second summer. In tropical areas the fish grow faster. To accelerate growth, warm-water ponds now exist in the temperate zone, where an average harvest of 400 to 500 kilograms per hectare is normal in intensive cultivation. By scientific management and careful selection it is possible to obtain yields up to 3,500 kilograms per hectare for carp in warm-water ponds.

Cold-water aquaculture

Trout. Although trout was the first fish to be artificially fertilized, trout cultivation in Europe and North America is much younger than carp cultivation. Trout are cold-water fish and must have a constant supply of sufficient oxygen, making cultivation more difficult. Though trout ponds can be smaller than carp ponds, good year-round water circulation is essential. Trout farms are therefore often located in mountainous areas where plentiful pure water is available. The young fish are obtained exclusively by artificial fertilization; thus, hatchery buildings with low-temperature water and good filters are the centre of this type of pond fishery. There the eggs are kept under control during breeding in special small tanks. As soon as the hatched fry can swim and eat on their own, they are transplanted to rearing ponds for feeding.

Trout are carnivores; meat-packing by-products are used for feed. Such food may be released into the ponds at predetermined intervals by automatic dispensers. Though many authorities claim that trout should have as much natural foodstuff as possible and therefore should be raised in natural ponds only, in many countries rearing is done in concrete-lined ponds or concrete tanks, which are easy to keep clean and permit disinfectant application. The time necessary to rear fish and the yield per hectare depend on feeding. Some trout farms sell their fish not only fresh and frozen but also smoked and filleted.

For trout and salmon, a new system of fish cultivation has been introduced. Instead of ponds, enclosures of netting or other materials are placed in natural waters, such as lakes, and also in brackish waters. By this means, areas formerly of low value can be farmed intensively. Farming trout in brackish water or seawater was of especial interest. Since the period preceding World War II, trout and salmon farming in seawater has grown tremendously.

Many other fish are raised artificially by various methods. Among these are sturgeon, milkfish, mullet, tilapias, striped bass, redfish, sea bass, and catfish.

Other types of aquaculture. *Mollusks.* Other important objects of cultivation in many parts of the world are mollusks. Though few water snails are cultivated, bivalves, especially oysters, are quite important in Asia, Europe, and North America. For centuries French fishermen cultivated oysters by placing twigs in the water to which free-swimming oyster larvae could attach. In northern Europe, oysters have been cultivated on the ocean bottom, but low winter temperatures limit the extent of this activity. In the Mediterranean, the Romans are said to have been the first to farm oysters.

Oyster cultivation

Today, oysters are cultivated on the Pacific coast of North America, as well as on the southern Atlantic coast and the Gulf of Mexico. Australia, the Philippines, and South Africa also possess farms, and the

Japanese grow edible oysters from Hokkaido in the north to Kyushu in the south. Japanese farms are divided into two classes: some cultivate seed oysters only, while others raise them for food, especially for export. The Japanese cultivate oysters on the sea bottom (horizontally) and on sticks (vertically). To collect the larvae, which affix themselves to any firm object, such as an old shell or a stone, fishermen place various devices in the water. These may be bamboo sticks with shells attached or a rope with shells hanging from it; limed tiles and wooden plates have been used for the same purpose in Europe. Production is greatest in places with good shelter against rough seas, a tidal current to carry food to the larvae, adequate salinity, and optimum temperature.

After some growing time, the larvae are loosened and transported to other areas for maturation under the best conditions. While growing to marketable size, the oysters must be protected against predators, such as starfish and oyster drillers. As starfish damage cannot be completely avoided when growing oysters on the bottom, a vertical system of culture is preferred in many areas; the oysters hang in clusters or in baskets or are fixed on poles in sheltered bays. In an alternative system, the oysters remain in horizontal trays kept at some distance from the bottom. Though such tray-raised oysters are expensive, they generally survive better than those reared directly on the bottom.

Blue mussels are cultivated in Italy, Spain, France, The Netherlands, and near Germany in the North Sea and the Baltic. There, too, horizontal-bottom methods have been replaced by vertical culture. Originally, the young mussels, collected from wild stocks, were spread on controlled banks leased by a fisherman from the government. Their capacity to grow in very extensive and dense beds is highly advantageous. Before full-grown mussels are sent in sacks to the market, special purification methods are employed to wash out sand. Today vertical culture is practiced with sticks pushed into the ocean bottom or with lines hanging from rafts. Unfortunately, line cultures may be damaged in winter; thus, experiments have been made with polyethylene net bags and endless tubes of polypropylene netting. These bags must be strong enough to carry the mussels until harvesting.

Many other mollusks are cultivated, including soft clams and scallops. The Japanese even raise octopuses and squid. For bivalves, the problems are roughly the same as mentioned above: collecting the larvae; raising the young mussels under good conditions; protecting them against predators; harvesting the adults without injury; and sometimes cleaning for the market.

Among inedible bivalves, pearl oysters deserve mention. Pearl farming is one of the most famous industries of Japan, dating to 1893, when a Japanese first succeeded in cultivating pearls. Under the skin of an oyster, the pearl farmer inserts a pearl nucleus (a small spherical shell fragment wrapped in a piece of living oyster tissue). The treated oyster is placed in a culture cage on a floating raft; after some months or years, the cultured oyster produces a pearl. Japan's pearl production is still concentrated along the coast of Mie Prefecture, where it was developed.

Pearl oyster culture

Crustaceans. Crustaceans—mainly shrimps, crayfish, and prawns—are also cultivated. In traditional Japanese practice, immature shrimps are caught in coastal waters and transferred to ponds. Today, mostly in the United States and Japan, shrimps are cultivated by catching adult egg-bearing females. The presence of eggs can be detected by examining the ovaries, usually visible through the shell. The female shrimps are transferred to large seawater ponds adjacent to the sea or to tanks. After hatching, the shrimps are fed in indoor tanks with cultivated plankton. After 10 days they are brought to shallow ponds for further cultivation or for distribution to other farms.

Americans, Australians, and Europeans have shown interest in commercial lobster culture. Production methods are not yet economical, however. The animals take two to three years to reach market size, and they have a high mortality rate. Lobsters must molt in order to grow and are quite vulnerable during the molting period.

Seaweed. A final important item in aquaculture is seaweed. Laver, a red alga, is a traditional part of the Japanese diet. The Japanese first cultivated this plant in the late 17th century in the brackish water of Tokyo Bay. Originally, a vertical method was used, with bushes placed in the water. A horizontal method is now employed: large meshed netting made of rough materials is hung horizontally between poles at the proper depth. The algae grow there by themselves and the owners harvest them from the nets by hand. Harvesting begins early in November and continues until about March. After the season is over, the gear is removed and stored. Part-time fishermen or land farmers often engage in such algae culture. Though other algae are used for food and in industry today, commercial farming has yet to begin. (A.v.B./C.H.A.)

WHALING

International whaling developed in stages that were determined by changing demand, diminishing stocks, and advancing technology. A lengthy primitive stage eventually led to commercial whaling; new markets and technical and chemical advances produced modern whaling; and virtual extinction of the quarry led back to a final primitive stage. The commercial stages were overwhelmingly dominated by northern Europeans and Americans—first by the Dutch, then by the British and Americans, and finally by the Norwegians and British. Only at the very end, when the Europeans found the trade no longer profitable, did they surrender the remaining whales to the Japanese.

Primitive whaling. Archaeological evidence suggests that primitive whaling, by Inuit and others in the North Atlantic and North Pacific, was practiced by 3000 BC and has continued in remote cultures to the present. The primitive quarry were small, easily beached whales or larger specimens that came close to shore during seasonal migrations from polar feeding grounds to breed in sheltered bays. The Japanese used nets, and the Aleuts used poisoned spears. The Inuit successfully hunted large whales from skin boats, employing toggle-head harpoons attached by hide ropes to inflated sealskin floats. A number of harpoons were darted into a whale, impeding the creature's escape until it could be safely killed with a lance. In Europe, the Nordic people hunted small whales, and Icelandic laws dealt with whaling in the 13th century.

Early whaling from boats

The forerunners of commercial whaling were the Basques, who caught black right whales as the animals gathered to breed in the Bay of Biscay. Docile and slow moving, sleeping on the surface, the whales were chased by rowboat, struck by harpoon, "played" like fish, and then lanced to death. Their bodies, which floated after death, were towed to shore for the stripping and boiling of their thick blubber and the processing of their baleen. When seaworthy oceangoing ships were built, Basques set off in search of other whaling bays and found them—perhaps in the 14th century and certainly by the 16th—across the Atlantic in southern Labrador.

Early commercial whaling. While the Basques acquired experience, northern Europe developed more capital and better markets. Drafting Basque whalemen for its Arctic explorations, the English Muscovy Company initiated the exploitation of whaling bays in Spitsbergen in 1610. The Dutch followed immediately, and, with a combination of violence and better business organization, broke the English monopoly, which had already stifled native competition. Smeerenburg ("Blubbertown") was built on Spitsbergen after 1619; in its heyday in the 1630s and '40s, it had 150 men servicing whalers that hunted the Greenland right, or bowhead, whale in the neighbouring Arctic.

The demise of Arctic bay whaling owed less to overfishing than to a miniature ice age lasting for the rest of the 17th century. Smeerenburg shut down in the 1660s, and Dutch and German whalers navigated the open sea ice. Whales were flensed (stripped) alongside the vessels, and their blubber (preserved to some extent by the cold) was brought home in barrels. This Greenland phase of whaling—extending into Davis Strait after 1719—was dominated by the Dutch and Germans until their activity collapsed in the 1780s. At that point Britain, in order to service its industrial revolution, took over as the principal European whaling nation.

Sperm whaling

Since the 1690s the British had "fished" from bay stations in their North American colonies, with Cape Cod, Long Island, and Rhode Island becoming the centres of activity. There a new type of whaling was inaugurated in 1712, when a Nantucket vessel caught the first sperm whale, whose waxy oil and spermaceti were worth far more than right whale oil. The sperm whale was a free-ranging oceanic animal, encouraging expeditions into warmer waters, where rapid putrefaction of blubber occurred. Fortunately, sperm whales had far less blubber than did right whales, and in the 1760s try-works (brick ovens for rendering blubber) appeared on American vessels. Henceforth, voyages were limited only by the whaler's capacity and the whaleman's endurance. Whale hunts extended from the Atlantic to the rich grounds of the Pacific, where four-year cruises were not uncommon, often mixing sperm whaling on the open sea with sealing and bay whaling for right whales.

During the great age of early commercial whaling in the late 18th and early 19th centuries, the chief means of capturing whales were relatively light double-ended rowing boats. A crew of only six men put off in pursuit with hand harpoons and a coiled line to play the whale, which was killed with a hand lance when it was sufficiently exhausted. American boats were usually nine metres long and made of half-inch cedar, while the British boats in Davis Strait were stronger.

Both northern and southern commercial whaling entered the doldrums around 1860. The U.S. fleet of more than 700 vessels (with 429 registered in New Bedford alone in 1857) declined rapidly, owing to overfishing but principally to the discovery, in Pennsylvania in 1859, of petroleum, which replaced sperm oil in the lamps and candles of America. The Dutch fleet had collapsed early in the century. The British Arctic fleet was devastated in the 1830s and '40s by overfishing and ice and by the introduction of vegetable oil, steel-boned corsets, and gas-fired lamps. Residual activity continued in the South Pacific and Davis Strait until the eve of World War I, and a late attack on North Pacific right whales from San Francisco ended in the 1920s.

Despite the growing potential of mineral and vegetable oils, there was still a healthy demand for whale, as opposed to sperm, oil. However, neither Britain nor America could provide a successful technological response to the problem of diminishing returns on their investments. Little financial benefit was gained from applying steam power to whalers, and iron vessels had inferior insulation properties and lacked the resilience to resist ice pressure in the Arctic. Harpoon guns and rocket harpoons were introduced after 1815, but they were largely abandoned as more likely to frighten off whales than guarantee their capture. "Darting guns," which fired grenades into harpooned whales, were little better in causing fatal hemorrhage than hand lances. Most seriously, with right whales on the point of extinction, neither country found a means of attacking the vast stock of alternative whales.

Steam and gun

Modern whaling. Tradition taught that only right and sperm whales were worth hunting. So-called wrong whales—also called rorquals—included the blue (the largest animal ever to have existed), the fin, and the sei. With a top speed of 20 to 30 miles per hour, rorquals were too fast for the six-oared whale boats and too heavy to be held by them when—in contrast to right whales—they sank after being killed. It was a Norwegian, Svend Foyn of Tønsberg, who brought the various strands of modern whaling together between 1863 and 1870. With his 86-ton, seven-knot *Spes et Fides,* the first powerful steam-powered whale catcher, he replaced the old rowing boats. A forward-mounted gun fired a heavy harpoon equipped with a time-delay grenade, and the main engines were used, via a winch and spring arrangement, to play, raise, and tow whales for processing ashore. For the first time, carcasses were fresh enough for the oil and meat to be edible. Modern cookers and machinery extracted oil from inedible meat and bones, and the residue was ground into fertilizer or animal feed. The extra income thus generated helped balance the poor baleen yield of rorquals.

Foyn's success led to other shore stations on the Norwegian, Scottish, and Newfoundland coasts by the turn of the century. Stocks were already diminishing when interest in whale oil suddenly increased around 1904 as the production of soap and margarine ran ahead of world fat supplies. The hydrogenation process, turning oil into fat, opened up whaling to immense capital investment. Norwegian and British enterprises established bay stations in the Antarctic, South America, and South Africa, and the catch rose from fewer than 2,000 whales per annum around 1900 to more than 20,000 from 1911. By 1909 whales were hunted in the open seas by fleets consisting of steam catchers and a factory ship that carried fuel and supplies and processed the whales. This pelagic whaling rose in importance after World War I, largely accounting for a rapid rise in oil production in the late 1920s. By 1930 there were six shore stations, 41 factory ships, and 232 catchers in the Antarctic. Stern slips, better butchering machinery, and pressure cookers speeded up processing and helped to raise the yield of good quality oil well ahead of the requirements of the manufacturing companies. Although much capital and commercial organization was provided by Britain, this was the great period of Norwegian domination of the trade, with many of the expeditions and almost all the crewmen of Norwegian origin.

After World War II (during which some 30 floating factories were sunk), so important was whale oil to the fat rations of Europe that a wave of newer, larger factories and more powerful 18-knot diesel-engined catchers was built, backed up by aircraft and support shipping. The success enjoyed by British and Norwegian companies (with over 80 percent of the trade from 1945 to 1950) brought in other nations—notably The Netherlands, the U.S.S.R., and Japan. The hunt descended from larger to smaller whales as overfishing decimated the blue whale in the 1940s, the fin and humpback whales between 1955 and 1970, and the sei whale thereafter. Whaling shifted from the Antarctic (where some 1,400,000 whales were killed) to the North Pacific. A renewed attack on the sperm whale took place in the 1950s, when new uses were found for its oil.

Modern catcher boats were about 60 metres long and designed for quick turning. From a raised foredeck a bow gun accurately fired a two-metre harpoon of 54 kilograms as far as 25 metres. The harpoon was made of soft metal, so that it bent, but did not break, when the animal struggled. A time fuse then triggered an explosive charge to damage vital organs or cause massive bleeding. A nylon line running from the harpoon to a power winch brought in the whale when it was exhausted; on occasion, a second shot was required to kill the animal. Once alongside, the carcass was inflated with air to keep it afloat, and an identifying flag was affixed. The catcher sought more prey, eventually collecting the entire day's catch for delivery to the factory.

Modern factory-ship processing

Whale processing by modern and floating factories was broadly standardized and highly efficient. Carcasses were winched onto ramps, where they were dismembered with speed to release body heat before excessive putrefaction ensued. Blubber was flensed lengthwise in three sections—much like peeling a banana—chopped, and quickly fed through deck hatches into high-pressure steam cookers. The bone, which was porous and oily, was also cut up and subjected to pressure cooking for oil extraction, and the residue was ground and bagged as bone meal. Flesh not used for human consumption was similarly processed for oil, and the remaining solids were kiln-dried and marketed as animal feed. In less than one hour it was possible to dispose of all that was useful of a 100-ton blue whale.

About 1962, with no more concentrations of large whales, the commercial expeditions from Europe withdrew and left whaling chiefly to the Soviets and Japanese. Soviet enterprises were subsidized, while modern refrigeration allowed the Japanese to make a delicacy out of the meat, which Europeans had used—if at all—for animal feed or fertilizer. But even this whaling was doomed to failure as it rapidly extinguished the last of the "commercial" whales, the minke. By the 1980s the international whale trade was dead, although small-scale whaling continued.

Regulation. The demise of whaling and the whale was regularly forecast from the 1930s, but restraint was hampered by the selfishness of commerce and the ignorance of science. In 1931 a League of Nations International Whaling Convention led the way into restrictive regulations, whose partial success in the 1930s owed more to poor business than good sense. The International Whaling Commission (IWC), established in 1946 to conserve whale stocks, prohibited the killing of right and gray whales. It also limited the annual Antarctic kill to 15,000 Blue Whale Units (BWU; one BWU = two fin, 2.5 humpback, or six sei), and it created closed areas and seasons around the world. Enforcement was difficult, however, and the BWU standard was overoptimistic about total and optimum whale stocks. Modern scientific pressure, coinciding with commercial crisis, brought the BWU limit down to 2,300 by 1962 and protected the humpback (1963) and blue (1965); but it was not until 1972, when the United Nations voted in vain for a moratorium on whaling, that the BWU was replaced by a New Management Procedure, which established quotas for the remaining species. In 1979 floating factories were banned, except for minke whales. Finally, to protect remaining stocks, the IWC agreed in 1982 to prohibit commercial whaling beginning in 1986, but several countries refused to agree. Residual commercial whaling continued, often under the guise of capturing specimens for scientific research.

Whaling moratorium

Whale products. *Whale oil.* This is a compound of fatty acids and glycerol, plus other alcohols and free fatty acids, the latter increasing with degeneration and producing rancidity. Premodern oil was inedible and was used principally for lighting, lubrication, and the manufacture of varnish, leather, linoleum, rough cloth (especially jute), and bottled gas for railway and similar uses. Until recently the finest lubricating oil came from the jawbones of small whales. Modern fresh oil, after being hardened to a fat, was used in margarine and soap; vegetable oil was not a practical alternative until the late 1930s. Whale liver oil was a major source of vitamin D through the 1960s.

Sperm oil. Waxy and inedible, sperm oil was a superior lighting oil. It was little used in the modern period apart from certain toiletries and pharmaceuticals. From 1950 more could be used in soap, owing to advances in oil chemistry—hence the last great attack on sperm whales.

Spermaceti. This liquid wax from the head of the sperm whale produced, on freezing and pressing, a brittle wax highly prized for candles.

Baleen. Also called whalebone, this product was cut from the large, hairy plates with which right whales filter krill during passive feeding. Because it is springy and retains shapes imposed with heat, it was used for corsets, knife handles, umbrella ribs, and brushes.

Sperm whale teeth. Widely used for scrimshaw, they are now copied in plastic.

Ambergris. A waxy pathological accretion occasionally found in the intestine of sperm whales or floating in the sea, ambergris was for many years highly prized as a base in perfumery.

Whale meat. This was of no value in early commercial whaling, because of rapid putrefaction. Modern shore and floating factories produced meat meal for animal consumption or fertilizer until refrigerated ships facilitated recovery of edible meat in the late 1940s. Most countries rejected it except as dog food, but its high value in Japan supported the last phase of commercial whaling. (G.J.)

BIBLIOGRAPHY

Commercial fishing: The productivity of the seas and the exploitation of the living resources in the world's oceans are illustrated in EOOD AND AGRICULTURE ORGANIZATION OE THE UNITED NATIONS, *Atlas of the Living Resources of the Seas,* 4th ed. (1981). ORGANISATION EOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Multilingual Dictionary of Fish and Fish Products,* 2nd ed. (1978, reprinted with corrections, 1984), provides comprehensive information on fish, fish products, and other marine life used commercially. Scientific research supporting commercial fishing is discussed in TAIVO LAEVASTU and MURRAY L. HAYES, *Fisheries Oceanography and Ecology* (1981); TONY J. PITCHER and PAUL J.B. HART, *Fisheries Ecology* (1982); and MAURICE E. STANSBY, *Industrial Fishery Technology,* 2nd ed. (1976). Developments in commercial fishing are detailed in studies of local areas, such as PETER POWNALL, *Fisheries of Australia* (1979); and PETER R. SINCLAIR, *State Intervention and the Newfoundland Fisheries* (1987). C.R.P.C. BRANSON (ed.), *Fishermen's Handbook,* 2nd ed. (1987), explains and illustrates many situations occurring in commercial fishing.

Equipment, facilities, and methods: JOHN C. SAINSBURY, *Commercial Fishing Methods: An Introduction to Vessels and Gears,* 2nd ed. (1986), describes the basics of the fishing industry. ANDRES VON BRANDT, *Fish Catching Methods of the World,* 3rd rev. ed. (1984), surveys fishing methods and equipment, from the most primitive to the totally computerized and automated. DAG PIKE, *Fishing Boats and Their Equipment* (1979), examines the design factors of boats and their effect on fishing productivity. JAN-OLOE TRAUNG (ed.), *Fishing Boats of the World,* 3 vol. (1955–67), provides excellent basic information on boats for various fishing methods. Specific methods and technologies are discussed in *Echo Sounding and Sonar for Fishing* (1980), a Food and Agriculture Organization fishing manual; DAVID B. THOMSON, *Seine Fishing: Bottom Fishing with Rope Warps and Wing Trawls,* rev. ed. (1981), and *Pair Trawling and Pair Seining: The Technology of Two-Boat Fishing* (1978); J.H. MERRITT, *Refrigeration on Fishing Vessels,* rev. ed. (1978); and J.J. CONNELL (ed.), *Advances in Fish Science and Technology* (1980). *Fishing Ports and Markets* (1970) is a collection of papers from a conference held under the auspices of the Food and Agriculture Organization.

Fisheries: The resources, ecology, and management of marine and freshwater fisheries are discussed in ROBERT J. BROWNING, *Fisheries of the North Pacific: History, Species, Gear, & Processes,* rev. ed. (1980); ROBERT T. LACKEY and LARRY A. NIELSEN (eds.), *Fisheries Management* (1980); FREDERICK W. BELL, *Food from the Sea: The Economics and Politics of Ocean Fisheries* (1978); STEPHEN CUNNINGHAM, MICHAEL R. DUNN, and DAVID WHITMARSH, *Fisheries Economics* (1985); G.C. EDDIE, *Engineering, Economics, and Fisheries Management* (1983); TOIVO LAEVASTU and HERBERT A. LARKINS, *Marine Fisheries Ecosystem: Its Quantitative Evaluation and Management* (1981); ROBIN L. WELCOMME, *Fisheries Ecology of Floodplain Rivers* (1979); ROBIN G. TEMPLETON (ed.), *Freshwater Fisheries Management* (1984); and SHELBY D. GERKING (ed.), *Ecology of Freshwater Fish Production* (1978). Ongoing research and developments in the field are reported in *Fisheries of the United States* (annual); and *Ocean Yearbook.* (J.C.Sa./D.Pi.)

Aquaculture: Various aspects of commercial fish farming throughout the world are discussed in ELISABETH MANN BORGESE, *Seafarm: The Story of Aquaculture* (1980); ROBERT KIRK, *A History of Marine Fish Culture in Europe and North America* (1987); MARCEL HUET, *Textbook of Fish Culture: Breeding and Cultivation of Fish,* 2nd ed. (1986; originally published in French, 1970); E. EVAN BROWN, *World Fish Farming: Cultivation and Economics,* 2nd ed. (1983); MALCOLM C.M. BEVERIDGE, *Cage Aquaculture* (1987); JAMES E. LANNAN, R. ONEAL SMITHERMAN, and GEORGE TCHOBANOGLOUS (eds.), *Principles and Practices of Pond Aquaculture* (1986); ROBERT R. STICKNEY, *Principles of Warmwater Aquaculture* (1979); FREDRICK W. WHEATON, *Aquacultural Engineering* (1977, reprinted 1985); JOHN P. STEVENSON, *Trout Farming Manual,* 2nd ed. (1987); ROBERT R. STICKNEY (ed.), *Culture of Nonsalmonid Freshwater Fishes* (1986); and EREDRICK W. WHEATON and THOMAS B. LAWSON, *Processing Aquatic Food Products* (1985). A good overview article is WILLIAM H. QUEEN, "Aquaculture: A Renewal of Interest," *Currents,* 3(3):20–25 (1988). (C.H.A.)

Whaling: Early descriptions of whaling include W. SCORESBY, *An Account of the Arctic Regions with a History and Description of the Northern Whale-Fishery,* 2 vol. (1820, reprinted 1969); and ALEXANDER STARBUCK, *History of the American Whale Fishery from Its Earliest Inception to the Year 1876* (1878, reissued with a new preface, 2 vol., 1964), featuring a list of all known voyages of 1715–1876. Early modern surveys, some of them with fascinating descriptions of whales, include J.T. JENKINS, *A History of the Whale Fisheries: From the Basque Fisheries of the Tenth Century to the Hunting of the Finner Whale at the Present Date* (1921, reprinted 1971); and R.B. ROBERTSON, *Of Whales and Men* (1954, reprinted 1969), a contemporary whaling narrative. Recent histories include LEONARD HARRISON MATTHEWS *et al., The Whale* (1968, reissued 1975), a general work on whaling; E.J. SLIJPER, *Whales,* trans. from Dutch, 2nd ed., edited by RICHARD J. HARRISON (1979); and HARRY MORTON, *The Whale's Wake* (1982).

The standard work on modern whaling remains J.N. TØNNESSEN and A.O. JOHNSEN, *The History of Modern Whaling* (1982; originally published in Norwegian, 4 vol., 1959–70). N.A. MACKINTOSH, *The Stocks of Whales* (1965); and K. RADWAY ALLEN, *Conservation and Management of Whales* (1980), examine the problem of extinction. Valuable information on whaling in the Arctic is found in *Arctic Whaling* (1984), one of a series of papers published by the Arctic Centre, University of Groningen, Neth. (G.J.)

Flatworms: Phylum Platyhelminthes

The platyhelminths, or flatworms, constitute the phylum Platyhelminthes, a group of invertebrate animals that are free-living as well as parasitic—*i.e.*, living on or in another organism and securing nourishment from it. They are bilaterally symmetrical (*i.e.*, the right and left sides are similar), usually flattened, and lack respiratory, skeletal, and circulatory systems; no body cavity (coelom) is present. The body is not divided into true segments; spongy connective tissue (mesenchyme) constitutes the so-called parenchyma and fills the space between organs. Flatworms, generally hermaphroditic—functional reproductive organs of both sexes occurring in one individual—are the lowest invertebrates to possess three embryonic layers—endoderm, mesoderm, and ectoderm—and to have a head region that contains concentrated sense organs and nervous tissue (brain).

The phylum consists of five classes: Trematoda (flukes), Cestoda (tapeworms), Turbellaria (planarians), Monogenea, and Aspidocotylea (or Aspidobothria). Members of all classes except Turbellaria are parasitic during all or part of the life cycle. Most turbellarians are exclusively free-living forms. Approximately 13,000 species of flatworms have been described.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 313, and the *Index*.

This article is divided into the following sections:

GENERAL FEATURES

Fluke and tapeworm infestations

Importance. Platyhelminthes are of particular economic interest because many species of flukes are parasitic in man, in domestic animals, or in both. In Europe and in North and South America, tapeworm infestations of man have been greatly reduced as a consequence of routine meat inspection. But where sanitation is poor and meat eaten undercooked, the incidence of tapeworm infestations is high. In the Baltic countries much of the population is infested with the broad tapeworm (*Diphyllobothrium latum*); in parts of the southern United States a small proportion of the population may be infested with the dwarf tapeworm (*Hymenolepis nana*). In Europe and the United States the beef tapeworm (*Taenia saginata*) is not uncommon due to the habit of eating undercooked steaks or other beef products.

Parasites in immature stages (larvae) can cause serious damage to the host. A larval stage of the gid parasite of sheep (*Multiceps multiceps*) usually lodges in the sheep brain. Fluid-filled hydatid cysts (*i.e.*, sacs containing many cells capable of developing into new individuals) of *Echinococcus* may occur almost anywhere in the body of sheep. Hydatids of the liver, brain, or lung of man are often fatal. Infestation occurs only where man lives in close association with dogs that have access to infested sheep for food.

Thirty-six or more species of flukes have been reported as parasitic in humans. Endemic (local) centres of infection occur in virtually all countries; widespread infections occur in the Far East, Africa, and tropical America. Many species are ingested as cysts, called metacercariae, in uncooked food—*e.g.*, the lung fluke *Paragonimus westermani* found in crayfish and crabs, the intestinal flukes *Heterophyes heterophyes* and *Metagonimus yokogawai* and the liver fluke *Opisthorchis sinensis* in fish, and the intestinal fluke *Fasciolopsis buski* on plants. Free-swimming larvae (cercariae) of blood flukes penetrate the human skin directly. In man these parasites and others listed below cause much misery and death. Control of certain flukes through the eradication of their mollusk hosts has been attempted but without much success.

Schistosomiasis (bilharziasis) is a major human disease caused by three species of the genus *Schistosoma*, known collectively as blood flukes. Africa and western Asia (*e.g.*, Iran, Iraq) are endemic centres for *S. haematobium; S. mansoni* also is found in these areas, as well as in the West Indies and South America. In the Far East, *S. japonicum* is the important blood fluke.

Among domestic animals, the sheep liver fluke (*Fasciola hepatica*) may cause debilitating and fatal epidemics (liver rot) in sheep. These animals become infected by eating metacercariae encysted on grass. Monogenea are common pests on fish in hatcheries and home aquariums (see also DISEASE: *Diseases of animals*).

Size range and diversity of structure. Most of the Turbellaria are less than five millimetres (0.2 inch) long, and many are microscopic in size. The largest of this class are the planarians, which may reach 0.5 metre (about 20 inches) in length. Trematoda are mostly from less than one to 10 millimetres (0.04 to 0.4 inch) long; some species, however, may grow to several centimetres. The smallest Cestoda are about one millimetre (0.04 inch) long, but some exceed 15 metres (50 feet) in length. The Monogenea range in length from 0.5 to 30 millimetres (0.02 to 1.2 inches). Aspidocotylea are from a few millimetres to 100 millimetres in length.

Distribution and abundance. In general, free-living flatworms can occur wherever there is moisture. Except for one group of turbellarians, the temnocephalids, flatworms are cosmopolitan in distribution. They occur in both fresh water and salt water and occasionally in terrestrial habitats that are moist, especially in tropical and subtropical regions. The temnocephalids, which are parasitic on freshwater crustaceans, occur primarily in Central and South America, Madagascar, New Zealand, Australia, and islands of the South Pacific.

Some flatworm species occupy a very wide range of habitats. One of the most cosmopolitan and most tolerant of different ecological conditions is the turbellarian *Gyratrix hermaphroditus*, which occurs in fresh water at elevations from sea level to 2,000 metres (6,500 feet) as well as in saltwater pools. Adult forms of parasitic flatworms are confined almost entirely to specific vertebrate hosts; the larval forms, however, occur in vertebrates and in invertebrates, especially in mollusks, arthropods (*e.g.*, crabs), and annelids (*e.g.*, marine worms). They are cosmopolitan in distribution, but their occurrence is closely related to that of the intermediate host or hosts.

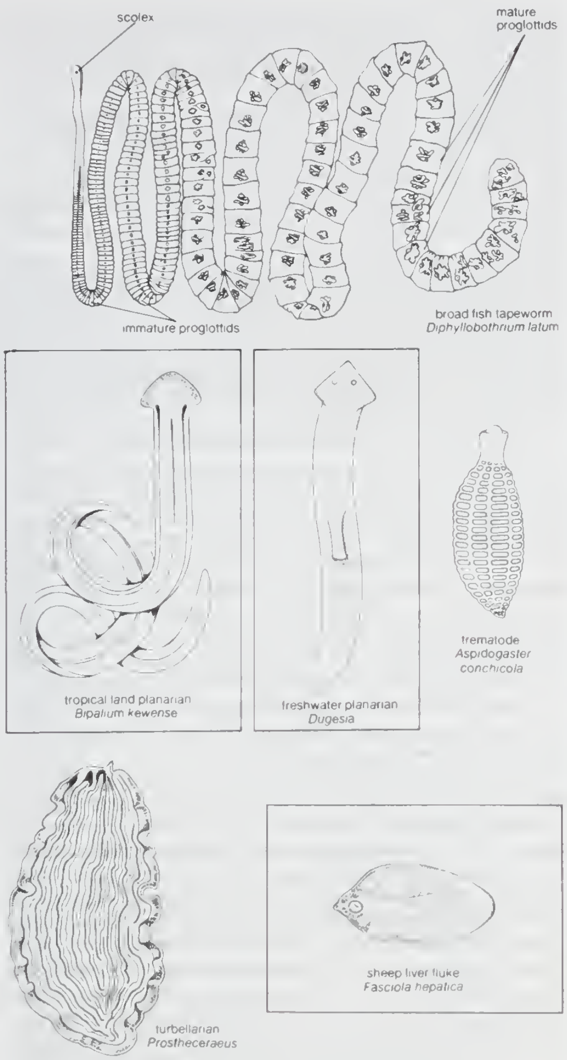


Figure 1: Representative body types of Platyhelminthes.

By courtesy of (bottom left) D.P. Wilson, from (top) Faust, Russell and Jung *Craig and Faust's Clinical Parasitology* (1970), reproduced by permission of Lea & Febiger (centre right) *Invertebrate Zoology* by Paul A. Meglitsch, Copyright © 1967 by Oxford University Press, Inc reprinted by permission (bottom right) © 1965 Time-Life Books, Time Inc

NATURAL HISTORY

Life cycle. *Reproduction.* With very few exceptions, platyhelminthes are hermaphroditic, and their reproductive systems are generally complex. Numerous testes but only one or two ovaries are usually present in these flatworms. The female system is unusual in that it is separated into two structures: the ovaries and the vitellaria, often known as the vitelline glands or yolk glands. In contrast, in most higher animals the yolk is part of the egg. The cells of the vitellaria form yolk and eggshell components. In some groups, particularly those that live primarily in water or have an aqueous phase in the life cycle, the eggshell consists of a hardened protein known as sclerotin, or tanned protein. Most of this protein comes from the vitellaria. In other groups, especially those that are primarily terrestrial or have a terrestrial phase in their life cycle, the eggshells are composed of another protein, keratin, a tougher material that is more resistant to adverse environmental conditions.

Separation of the female reproductive system

The general structure of the reproductive system may be seen in Figure 2. In the tapeworms, the tapelike body is generally divided into a series of segments, or proglottids, each of which develops a complete set of male and female genitalia. A rather complex copulatory apparatus consists of an evertible (capable of turning outward) penis, or cirrus, in the male and a canal, or vagina, in the female. Near its opening the female canal may differentiate into a variety of tubular organs. Fertilized eggs are often stored in a saclike uterus, which may become greatly distended; in tapeworms, it may fill a whole segment.

Each male and female reproductive system may have its own external opening, or gonopore, or the terminal regions of each system may join to form a common genital atrium (Figure 2), or passage, and a genital pore.

Either cross-fertilization (*i.e.,* involving two individuals) or self-fertilization may occur; self-fertilization is probably more common. Some free-living flatworms perform a type of copulation known as hypodermic impregnation, whereby the penis of one animal pierces the epidermis of another and injects sperm into the tissues. Some forms reproduce asexually—*i.e.,* they produce chains that eventually separate to form individuals.

Modes of fertilization

Development. The life cycles of the free-living forms are relatively simple. Eggs are laid singly or in batches. Frequently they are attached to some object or surface by an adhesive secretion. After a period of embryonation (*i.e.,* the formation of the first larval stage) minute worms emerge, feed, develop genitalia, and become adult.

Parasitic platyhelminths undergo very complex life cycles, often involving several larval stages in other animals—the intermediate hosts; these hosts may be invertebrate or vertebrate.

The simplest cycle in parasitic platyhelminths occurs in the Monogenea, which have no intermediate hosts. The majority of the Monogenea are ectoparasitic (externally parasitic) on fish. The eggs hatch in water. The larva, known as an oncomiracidium, is heavily ciliated (has hairlike projections) and bears numerous posterior hooks. It attaches to a host, grows, and matures. In some species (*e.g., Polystoma integerrimum*) parasitic in frogs, maturation of the genitalia is synchronized with maturation of the host and apparently is controlled by the endocrine system of the latter.

In the life cycle of flukes of the subclass Digenea, mollusks serve as the intermediate host. Eggs of Digenea usually hatch in water or in a snail host. The first larval stage, the miracidium, generally is free-swimming and penetrates a freshwater or marine snail, unless it has already been ingested by one. Within this intermediate host, the parasite passes through some or all of a series of further larval stages known as sporocysts, rediae, and cercariae. Cercariae are essentially young flukes with tails. They either penetrate the host or are ingested by it. Tapeworms of the subclass Eucestoda are generally transmitted from host to host by direct ingestion of eggs; by ingestion of intermediate hosts containing larval stages; and, very rarely, by passage of a larva from an intermediate host through a skin wound into another intermediate host.

Transmission to a human host through a skin wound is most likely to occur in Asia, where frogs infested with tapeworm larvae are sometimes used to treat wounds. The tapeworm, *Hymenolepis nana,* parasitic in rodents and man, can complete its life cycle without an intermediate host. Certain species of trematodes and cestodes show a tendency toward progenesis, in which adult features—such as the appearance of genital rudiments—appear in the larva. In some cases of progenesis the worm achieves sexual maturity in the intermediate host.

Regeneration. The ability to undergo tissue regeneration, beyond simple wound healing, occurs in two classes of Platyhelminthes: Turbellaria and Cestoda.

Turbellaria. Turbellarians, planaria particularly, have been commonly used in regeneration research. The greatest regenerative powers exist in species capable of asexual reproduction. Pieces from almost any part of the turbellarian *Stenostomum,* for example, can reorganize into completely new worms. In some cases regeneration of very small pieces may result in the formation of imperfect (*e.g.,* headless) organisms.

In other Turbellaria, regeneration of the head is limited to pieces from the anterior region or to tissues containing the cerebral ganglion (brain). The region anterior to this ganglion is incapable of regeneration, but if cuts are made posterior to it, many species can replace the entire posterior region, including the pharynx and the reproductive system. In the cut pieces, polarity is retained; *i.e.,* the anterior zone of the cut piece regenerates the head and the posterior region regenerates the tail. If a region in front of the pharynx is transplanted into the posterior region

Polarity

of another individual, it influences that region to form a pharyngeal zone that eventually differentiates a pharynx. This new pharyngeal zone is now said to be determined and, if removed, will regenerate again into a new pharynx.

There is evidence that a special type of cell, a neoblast, is concerned in planarian regeneration. Neoblasts, rich in ribonucleic acid (RNA), which plays an essential role in cell division, appear in great numbers during regeneration. Similar cells, apparently inactive, occur in the tissues of whole organisms (see also GROWTH AND DEVELOPMENT: *Biological regeneration*).

Cestoda. Regeneration, although rare in the parasitic worms in general, does occur in the cestodes. Most tapeworms appear to be capable of regeneration from the head (scolex) and neck region. This property often makes it difficult to treat man for tapeworms; treatment may eliminate only the body, or strobila, leaving the scolex still attached to the intestinal wall of the host and thus capable of producing a new strobila, which reestablishes the infestation. Several species of cestode larvae can regenerate themselves from cut regions. A branched larval form of *Sparganum prolifer,* a human parasite, may undergo both asexual multiplication and regeneration.

Ecology. Turbellaria are adapted to a wide range of environments, and many species are resistant to extreme environmental conditions. Some occur in coastal marine habitats—in sand, on or under rocks, and in or on other animals or plants. Some marine species occur at relatively great depths in the sea; others are pelagic (*i.e.,* living in the open sea). Freshwater species are found in ponds, lakes, rapidly flowing rivers, and streams. Temporary freshwater pools may contain adult forms that survive periods of dryness in an encysted state. Some aquatic species exhibit considerable tolerance to osmotic changes—*i.e.,* to differences in salt concentrations of the water; a marine species *Coelogynopora biarmata,* for example, has also been found in freshwater springs.

Terrestrial species of Turbellaria occur in soil, moist sand, leaf litter, mud, under rocks, and on vegetation. Some have been found in pools in the desert and in caves. Cave-dwelling species tend to show loss of eyes and pigment.

Tolerance of temperature extremes

Some species are able to stand considerable temperature ranges; for example, *Crenobia alpina,* which occurs in alpine streams, apparently can survive temperatures of −40° to −50° C (−40° to −58° F). Remarkable heat tolerance is exhibited by *Macrostomum thermale* and *Microstomum lineare,* which are found in hot springs at 40°–47° C (104°–117° F). *M. lineare* can also tolerate temperatures as low as 3° C (37° F).

Many Turbellaria live in association with plants and animals. Marine algae, for example, frequently harbour many species, often in large numbers. Turbellarians most commonly associate with animals such as echinoderms (*e.g.,* starfish), crustaceans (*e.g.,* crabs), and mollusks. Less commonly, associations occur with sipunculids (tiny marine invertebrates), annelid worms, arachnids (*e.g.,* spiders), coelenterates (*e.g.,* jellyfish), other turbellarians, and lower vertebrates. An interesting feature of these associations is that species within a turbellarian family tend to associate with one type of organism; for example, almost all members of the family Umagillidae associate with echinoderms.

In a few cases, the association is considered parasitic; *i.e.,* the turbellarians obtain all nourishment from the host. Most of these species belong to the order Rhabdocoela, in which the alimentary canal is either absent or reduced. In a few hosts they cause so-called parasitic castration—*i.e.,* suppression of gonad maturation in the host organism.

Among the turbellaria that are parasitic or commensal (*i.e.,* living in close association with but not harmful to another organism) the Temnocephalida are best adapted for attachment to other organisms. They have a large saucer-shaped posterior adhesive organ and anterior tentacles that are also used for adhesion. All temnocephalids occur on freshwater hosts, mainly crustaceans but also mollusks, turtles, and jellyfish.

The tendency to associate with other animals apparently represents a definite evolutionary trend among the platyhelminths; permanent associations essential to the survival of a species could develop from loose associations, which could also give rise to parasitic forms, such as the Trematoda and Cestoda. The free-living larval stages that frequently occur in these groups play a major role in the dissemination of the species.

The ecology of the parasitic groups (*i.e.,* Cestoda and Trematoda) is particularly complex, because as many as four hosts may be involved in the life cycle. In the case of the broad tapeworm, for example, man serves as the final (or definitive) host, various species of fish as one intermediate host, and species of a small water crustacean (*Cyclops*) as another intermediate host. It is clear that the broad tapeworm (*Diphyllobothrium latum*) can occur only where an intimate ecological association exists among the three host groups.

Adaptation to the microenvironment of the host

In addition to adapting to the general external environment, individuals at each stage of the life cycle of a parasite must adapt to the microenvironment provided by the host. Adaptations include not only obvious features, such as suckers or hooks for attachment, but also those associated with the biochemical and physiological conditions imposed by the host. Parasites frequently utilize the physiological and biochemical properties of a new host, especially those that differ markedly from the external environment, in order to trigger the next developmental stage—*e.g.,* several species of cestodes are stimulated to mature sexually by the high body temperature (40° C) of their bird host, which contrasts sharply with the low body temperature of the cold-blooded fish host of the larval stage. The unusually intimate association of certain flukes (subclass Digenea) with mollusks suggests that flukes were originally parasites of mollusks and that they later developed an association with other hosts.

Knowledge of the ecology both of a platyhelminth parasite and of its intermediate host(s) is essential if control measures against the pest are to be effective. Man's alterations of the environment sometimes increase the occurrence of platyhelminth diseases. The Aswan High Dam in Egypt, for example, has produced conditions especially favourable for the breeding of the snail that serves as the intermediate host of the blood fluke (*Schistosoma mansoni*). In this case, as with many trematode infestations, man exposes himself to the disease by bathing in water containing infective larvae (cercariae) released from infested snails; the cercariae enter directly through the skin. Certain other human diseases of platyhelminth origin—such as hydatid (cyst) disease, caused by the tapeworm *Echinococcus granulosus*—owe their survival and dissemination to man's close ecological association with dogs.

In the parasitic platyhelminth groups (*e.g.,* the Monogenea) that do not normally utilize intermediate hosts, there is a close ecological association between egg release and production of young of both the parasite and its host; infection of the next generation of host could not otherwise occur.

Many platyhelminths show highly specific adaptations to internal host environments. Many monogeneans, for example, show a marked preference for a particular gill arch in a fish. The scolex (head) of certain tapeworms of elasmobranch fishes (*e.g.,* sharks, skates, and rays) is highly specialized and can satisfactorily attach only to the gut of a fish possessing a complementary structure.

FORM AND FUNCTION

External features. Some Turbellaria are gray, brown, or black, with mottled or striped patterns. Others, which contain algae in the mesenchyme, are green or brown. Parasitic flatworms usually have no pigment, but cestodes may be coloured by food (*e.g.,* bile, blood) in their gut. Some parasitic forms may show masses of dark eggs through a translucent, creamy, or whitish tissue.

The flatworm body plan

The typical flatworm body is flattened and leaflike or tapelike (Figure 2). The head may be set off from the body or grade imperceptibly into it. The anterior (head) end can usually be distinguished from the posterior end in free-living forms by the presence of two pigment spots, which are primitive eyes. In the case of the tapeworm, the scolex is usually conspicuous for its breadth. The strobila (body) typically consists of proglottids, each of which is usually a self-sufficient reproducing unit with all of the sexual

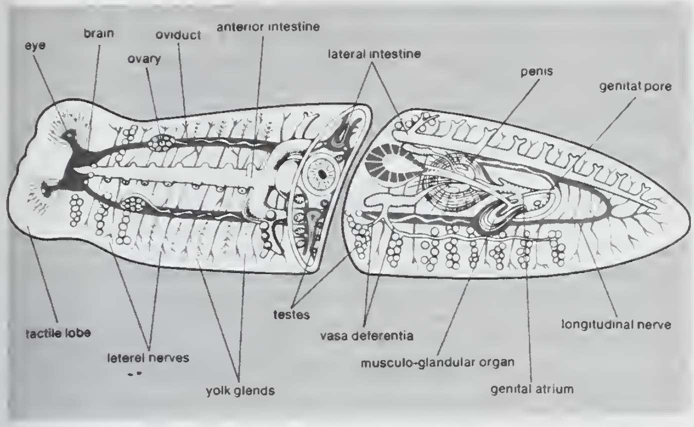


Figure 2: General structure of a platyhelminth, as exemplified by a freshwater turbellarian, *Procotyla fluviatilis*.

Reprinted with permission of The Macmillan Company from *Invertebrate Zoology* by R Hegner and J Engemann Copyright © by The Macmillan Company, 1968

organs necessary to reproduce. The number of proglottids may vary from three in some species to several hundreds in others. Organs of attachment on the scolex may, in addition to suckers, consist of hooks, spines, or various combinations of these.

The structure and function of the body covering, or tegument, differs markedly between free-living and parasitic forms. In free-living forms, the body covering is typically an epidermis consisting of one layer of ciliated cells—*i.e.,* cells with hairlike structures—the cilia being confined to specific regions in some species. In the parasitic groups—flukes, tapeworms, and monogeneans—the tegument shows striking modifications associated with the parasitic way of life. It once was thought that the tegument is a nonliving secreted layer; it is now known, however, that the tegument of parasites is (Figure 3) metabolically active and consists of cells not separated from one another by cell walls (*i.e.,* a syncytium). The tegument itself consists of cytoplasmic extensions of tegumental cells, the main bodies of which lie in what may be described as the "subcuticular" zone, although a true cuticle is not present. A membrane separates the inner zone of the tegumental cells, the so-called perinuclear cytoplasm, from the surface syncytium, or distal cytoplasm (Figure 3).

The surface of tapeworms and monogeneans is drawn out into spinelike structures called microtriches, or microvilli. The microtriches probably help to attach the parasite to the gut of the host, absorb nutritive materials, and secrete various substances. In the flukes, microtriches are lacking, but spines are frequently present.

From (Left) R Hegner and J Engemann, *Invertebrate Zoology*, (Right) J Smyth, *The Physiology of Cestodes*

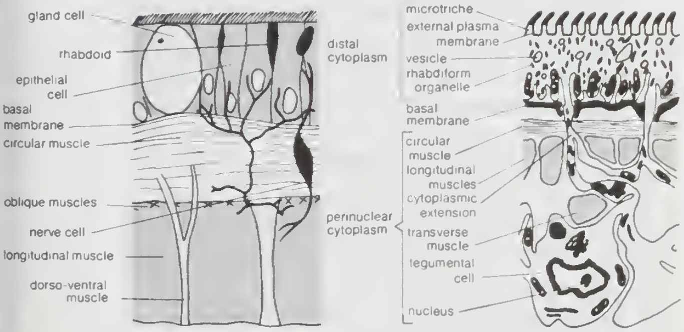


Figure 3: Comparison of the body wall in (left) a turbellarian, a free-living platyhelminth, and in (right) a tapeworm, a parasitic platyhelminth.

Embedded in the epidermis of turbellarians are ovoid or rod-shaped bodies (rhabdoids, see Figure 3) of several sorts; of uncertain function, the bodies frequently are concentrated dorsally or may be clustered anteriorly as rod tracts opening at the apex. Rhabdoids are absent in flukes and tapeworms.

Internal features. Beneath the epidermis of turbellarians is a homogeneous or lamellated basal membrane. Club-shaped mesenchymal gland cells, opening externally, generally are present in all flatworms. In turbellarians two major types of mesenchymal glands occur: one produces a slimy material upon which the organisms creep; the other secretes an adhesive substance for capture of prey, for adhesion, and for cementing egg capsules to a suitable surface. The larvae of parasitic forms generally possess similar glands whose secretions are used for adhesion, for producing cyst walls around resting stages, and for penetrating hosts; some adult parasites have glands for adhesion and, in trematodes, for softening and digesting host tissues.

Mesenchyme

The mesenchyme consists of fixed and free cells as well as a fibrous matrix. A fluid occupies the minute open spaces and serves for distribution of nutrients and wastes. The mesenchymal cells in certain groups may differentiate during growth to become sex cells or may function in asexual reproduction in repair or in regeneration.

Nervous system. The main ganglia, or nerve centres, of the nervous system and the major sense organs are generally concentrated at the anterior end. Typically, the primitive brain of the flatworm consists of a bilobed mass of tissue with lateral longitudinal nerve cords connected by transverse connectives, thus forming a rather ladderlike structure or grid running the greater length of the organism (Figure 2). Free-living forms commonly have two longitudinal cords, but some tapeworms have as many as 10. Sensory receptors occur in all groups.

Musculature. The well-developed muscular system present in flatworms is comprised of a subcuticular musculature consisting of layers of circular, longitudinal, and diagonal muscles close to the epidermis, and a mesenchymal musculature consisting of dorsoventral, transverse, and longitudinal fibres passing through the mesenchyme. In general, platyhelminths are capable of extensive body contraction and elongation.

Digestive and excretory systems. The blind-ending intestine of trematodes consists of a simple sac with an anterior or midventral mouth or a two-branched gut with an anterior mouth; an anus is usually lacking, but a few species have one or two anal pores. Between the mouth and the intestine are often a pharynx and an esophagus receiving secretions from glands therein. The intestine proper, lined with digestive and absorptive cells, is surrounded by a thin layer of muscles that effect peristalsis; *i.e.,* they contract in a wavelike fashion, forcing material down the length of the intestine. In many larger flukes lateral intestinal branches, or diverticula, bring food close to all internal tissues. Undigested residue passes back out of the mouth.

The digestive system of Turbellaria typically consists of mouth, pharynx, and intestine. In the order Acoela, however, only a mouth is present; food passes directly from the mouth into the parenchyma, to be absorbed by the mesenchymal cells. Tapeworms lack digestive systems.

The excretory system consists of protonephridia. These are branching canals ending in so-called flame cells—hollow cells with bundles of constantly moving cilia.

Complex feeding mechanisms

Nutrition. *Free-living forms.* Free-living platyhelminths (Turbellaria), mostly carnivorous, are particularly adapted for the capture of prey. Their encounters with prey appear to be largely fortuitous, except in some species that release ensnaring mucus threads. Because they have developed various complex feeding mechanisms, most turbellarians are able to feed on organisms much larger than themselves, such as annelids, arthropods, mollusks, and tunicates (*e.g.,* sea squirts). In general, the feeding mechanism involves the pharynx which, in the most highly developed forms, is a powerful muscular organ that can be protruded through the mouth. Flatworms with a simple ciliated pharynx are restricted to feeding on small organisms such as protozoans and rotifers, but those with a muscular pharynx can turn it outward, thrust it through the tegument of annelids and crustaceans, and draw out their internal body organs and fluids. Turbellaria, with a more advanced type of pharynx, can extend it over the captured prey until the animal is completely enveloped.

Digestion is both extracellular and intracellular. Digestive enzymes (biological catalysts), which mix with the food in the gut, reduce the size of the food particles. This partly digested material is then engulfed (phagocytized) by cells or absorbed; digestion is then completed within the gut cells.

Parasitic forms. In the parasitic groups with a gut (Trematoda and Monogenea), both extracellular and intracellular digestion occur. The extent to which these processes take place depends on the nature of the food. When fragments of the host's food or tissues other than fluids or semifluids (*e.g.,* blood and mucus) are taken as nutrients by the parasite, digestion appears to be largely extracellular. In those that feed on blood, digestion is largely intracellular, often resulting in the deposition of hematin, an insoluble pigment formed from the breakdown of hemoglobin. This pigment is eventually extruded by disintegrating gut cells.

Despite the presence of a gut, trematodes seem able to absorb certain materials, such as glucose, through the metabolically active tegument covering the body surface. Tapeworms, which have no gut, absorb all food through the tegument. The entry of amino acids (the structural units of proteins) and small molecules of carbohydrate (*e.g.,* sugars) through the tegument occurs by a mechanism called active transport, in which molecules are taken up against a concentration gradient. This process, similar to that in the vertebrate gut, requires the expenditure of energy. There is also evidence that cestodes may be able to digest materials in contact with the tegument by means of so-called membrane digestion, a little-understood process.

Metabolism. Both free-living and parasitic platyhelminths utilize oxygen when it is available. Most of the parasitic platyhelminths studied have a predominantly anaerobic metabolism (*i.e.,* not dependent upon oxygen). This is true even in species found in habitats—such as the bloodstream—where oxygen is normally available.

Parasitic platyhelminths are made up of the usual tissue constituents—protein, carbohydrates, and lipids—but, compared to other invertebrates, the proportions differ somewhat; *i.e.,* the carbohydrate content tends to be relatively high and the protein content relatively low. In larval and adult cestodes, carbohydrate occurs chiefly as animal starch, or glycogen, which acts as the main source of energy for species in low oxygen habitats. The level of glycogen, like other chemical constituents, can fluctuate considerably, depending on the diet or feeding habits of the host. In some species, more than 40 percent of the worm's dried weight is glycogen.

Because carbohydrate metabolism is important in parasitic flatworms, a substantial amount of carbohydrate must be present in the host diet to assure normal growth of the parasite. Hence the growth rate of the rat tapeworm (*Hymenolepis diminuta*) is a good indicator of the quantity of carbohydrate ingested by the rat. Experiments have shown that most parasitic worms have the capability of utilizing only certain types of carbohydrate. All tapeworms that have been studied thus far utilize the sugar glucose. Many tapeworms can also utilize galactose, but only a few can utilize maltose or sucrose.

An unusual constituent of both trematodes and cestodes is a round or oval structure called a calcareous corpuscle; large numbers of them occur in the tissues of both adults and larvae. Their function has not yet been established, but it is believed that they may act as reserves for such substances as calcium, magnesium, and phosphorus.

The chief proteins in cestodes and trematodes are keratin and sclerotin. Keratin forms the hooks and part of the protective layers of the cestode egg and the cyst wall of certain immature stages of trematodes. Sclerotin occurs in both cestode and trematode eggshells, especially in those that have larval stages associated with aquatic environments.

Requirements for egg hatching

Platyhelminth eggs hatch in response to a variety of different stimuli in different hosts. Most trematode eggs require oxygen in order to form the first larval stages and light in order to hatch. Light is thought to stimulate the release of an enzyme that attacks a cement holding the lid (operculum) of the egg in place. A similar mechanism probably operates in cestodes (largely of the order Pseudophyllidea) whose life cycles involve aquatic intermediate hosts or definitive hosts, such as birds or fish.

In many cestodes, especially those belonging to the order Cyclophyllidea, the eggs hatch only when they are ingested by the host. When the host is an insect, hatching sometimes is apparently purely a mechanical process, the shell being broken by the insect's mouthparts. In vertebrate intermediate hosts, destruction of the shell depends largely on the action of the host's enzymes. Activation of the embryo within the shell and its subsequent release depend on other factors, including the amount of carbon dioxide present, in addition to the host's enzymes. Factors involving a vertebrate host are also important in establishing trematode or cestode infections after encysted or encapsulated larval stages have been ingested. Under the influence of the same factors, tapeworm larvae are stimulated to evaginate their heads (*i.e.,* turn them inside out, so to speak), a process that makes possible their attachment to the gut lining.

EVOLUTION

Origin of Turbellaria

The origin of the platyhelminths and the evolution of the various classes have not been positively settled. There are, however, two main lines of thought. According to the more widely accepted view, the Turbellaria and an aberrant group, the Ctenophora, have been derived independently from the lower Coelenterata. In the alternative view, the Turbellaria are derived from the Ciliata and the Ctenophora from the Turbellaria.

It is generally believed that the parasitic groups are derived from the Turbellaria, many of which form close associations with other animals. These associations often show host specificity, a characteristic of truly parasitic forms. There are a number of views regarding the evolutionary relationships among the various parasitic groups. One school of thought proposes that rhabdocoel turbellarians gave rise to monogeneans; these, in turn, gave rise to digeneans, from which the cestodes were derived. Another view is that the rhabdocoel ancestor gave rise to two lines; one gave rise to monogeneans, who gave rise to digeneans, and the other line gave rise to cestodes. A further modification of the latter view, based largely on the study of the larval forms, proposes that cestodes were derived from monogeneans.

In considering the evolution of the parasitic groups, the digeneans should be mentioned in particular. The life cycle of digeneans is dominated by mollusks, which invariably act as intermediate hosts. This condition has led to the widely accepted view that digeneans were originally parasites of mollusks and later formed an association with vertebrates; the vertebrates, in turn, became incorporated into the life cycle as definitive hosts.

CLASSIFICATION

Distinguishing taxonomic features. The phylum Platyhelminthes is grouped with several other phyla (Nemertea, Rotifera, Nematoda, Gastrotricha, Acanthocephala, and Nematomorpha) as acoelomate triploblasts—*i.e.,* lacking a body cavity and with three embryonic layers: endoderm, mesoderm, and ectoderm. These phyla also represent animals that are unsegmented and whose body space tends to be occupied by mesenchyme. In Platyhelminthes, mesenchyme cells completely fill this space.

In the classification of platyhelminths the principal criteria are: habitat of organism (*i.e.,* free-living or parasitic); the characteristics of the body covering; the form and position of organs for attachment to host (when present); the presence or absence of segmentation; the form of the reproductive system, especially with respect to vitellaria (yolk glands) and the number of testes; the presence or absence of an alimentary canal; the characteristics of the pharynx (when present); and the nature of protective egg membranes.

Annotated classification. There is no unanimity concerning the classification of platyhelminths. The following classification is based partly on that of Grassé, Hyman, Yamaguti, and La Rue (see below *Bibliography*).

PHYLUM PLATYHELMINTHES (flatworms)

Flat, unsegmented worms; gastrovascular cavity and respiratory, skeletal, and circulatory systems absent; excretion by means of flame-bulb protonephridia; mesenchyme fills all spaces between organ systems; a variable number of longitudinal nerve cords with transverse connectives; body structure triploblastic (*i.e.,* 3 embryonic layers); reproductive system hermaphroditic and complex.

Class Turbellaria

Epidermis usually ciliated at least in part, provided with rhabdoids (minute rodlike structures); body unsegmented; gut present except in order Acoela; life cycle simple; mostly free-living, some ectocommensal, endocommensal (*i.e.,* living, respectively, outside or inside another organism without harming it), or parasitic; about 3,000 species.

Order Acoela. Exclusively marine; mouth present; pharynx simple or lacking; no intestine; without protonephridia, oviducts, yolk glands, or definitely delimited gonads.

Order Rhabdocoela. Saclike intestine; protonephridia and oviducts usually present; gonads few, mostly compact; nervous system generally with 2 longitudinal trunks.

Order Alloecoela. Pharynx simple, bulbose or plicate (many ridges); intestine may have short diverticula, or pockets; protonephridia paired; testes usually numerous; penis papilla generally present; nervous system with 3–4 trunks.

Class Monogenea

Oral sucker lacking or weakly developed; posterior end with large posterior adhesive disk (opisthaptor) usually provided with hooks; excretory pores paired, anterior and dorsal; parasites of the skin and other superficial locations, especially on the gills of fish; life cycle simple, no alternation of hosts; about 1,350 species.

Class Aspidocotylea (Aspidobothria)

Oral sucker absent; main adhesive organ occupying almost the entire ventral surface and consisting of suckerlets arranged in rows; excretory pore single and posterior; endoparasites of vertebrates, mollusks, and crustaceans; about 100 species.

Class Cestoda (tapeworms)

Elongated endoparasites with alimentary canal lacking; epidermis modified for absorption and secretion; usually divided into segments (proglottids); adhesive organs limited to anterior end; except in Cestodaria, adult stages almost entirely parasites of vertebrates; life cycles complicated with 1 or more intermediate hosts; about 3,500 species.

Subclass Cestodaria

Unsegmented tapeworms containing 1 set of genitalia; parasites of the body cavity or intestine of annelid worms or fish.

Order Amphilinidea. Uterus long and N-shaped; genital pores at or near posterior extremity; intestinal parasites of teleosts (bony fish).

Order Caryophyllidea. Uterus a coiled tube; genital pore well separated from posterior extremity; intestinal parasites of teleosts, occasionally in annelids.

Order Gyrocotylidea. Testes confined to anterior region; genital pores near anterior end; parasitic in intestine of fish of the genus *Chimaera.*

Subclass Eucestoda

Polyzoic tapeworms with scolex (head) of varying structure; body usually with distinct external segmentation; parasitic in intestine of vertebrates.

Order Tetraphyllidea. Scolex with 4 bothridia (leaflike muscular structure); vitellaria located in lateral margins of proglottids; genital pores lateral; parasites of elasmobranchs.

Order Lecanicephalidea. Reproductive system similar to Tetraphyllidea, but scolex divided into an upper disklike or globular part and a lower collarlike part bearing 4 suckers; mainly parasites of elasmobranchs.

Order Proteocephalidea. Scolex with 4 suckers, sometimes a 5th terminal one; vitellaria located in lateral margins; genital pores lateral; mainly parasites of cold-blooded vertebrates.

Order Diphyllidea. Two bothridia, each sometimes bisected by a median longitudinal ridge; large rostellum (cone-shaped or cylindrical projection) armed with dorsal and ventral groups of large hooks; cephalic peduncle (fleshy stalk on head) with longitudinal rows of T-shaped hooks; genital pore median, parasitic in elasmobranchs; 1 genus, *Echinobothrium.*

Order Trypanorhyncha. Scolex with 2 or 4 bothridia; vitellaria in continuous sleevelike distribution; parasites of elasmobranchs.

Order Pseudophyllidea. Scolex with 2 elongated, shallow bothria, 1 dorsal and 1 ventral; genital pore lateral or median. Vitellaria lateral or extending across proglottid and encircling other organs; parasites of teleosts and land vertebrates.

Order Nippotaeniidea. Scolex bears 1 apical sucker; parasites of freshwater fish; 1 genus, *Nippotaenia.*

Order Cyclophyllidea (Taenoidea). Scolex with 4 suckers; no uterine pores; 1 compact vitellarium behind ovary; mainly parasites of birds and mammals.

Order Aporidea. No sex ducts or genital openings; parasites of swans.

Order Spathebothriidea. Scolex without true bothria or suckers; strobila with internal segmentation but no external segmentation; parasites of marine teleosts.

Class Trematoda (flukes)

Ectoparasites or endoparasites; no ciliated epidermis; body undivided; adhesive organs well-developed; life cycles generally complex with 2 or more hosts, about 6,250 species.

Subclass Aspidogastrea

Oral sucker absent; main adhesive organ occupying almost the entire ventral surface, consists of suckerlets arranged in rows; excretory pore single and posterior; endoparasites of vertebrates, mollusks, and crustaceans.

Subclass Digenea

Oral and ventral suckers generally well-developed; development involves at least 1 intermediate host; usually endoparasites of vertebrates.

Superorder Anepitheliocystidia. Excretory bladder thin walled (*i.e.,* not epithelial in any stage); excretory pores open to exterior in tail; stylet (piercing structure) always lacking.

Order Strigeatoidea. Cercaria (immature form) fork-tailed; penetration glands present; 1–2 pairs of protonephridia.

Order Echinostomida. Cercaria with simple tail and many cyst-producing glands; life cycle with 3 hosts.

Superorder Epitheliocystidia. Cercaria completely lacking caudal excretory vessels at any stage of development; stylet present or lacking.

Order Plagiorchida. Cercaria typically armed with a stylet; encystment in invertebrates, rarely vertebrates; excretory vessels not open to the exterior.

Order Opisthorchiida. Cercaria never armed; excretory pores open on margins of tail.

Critical appraisal. There is disagreement on many aspects of the taxonomy of Platyhelminthes, especially regarding class divisions. Among the free-living or semiparasitic forms, the temnocephalids are placed as an order among the class Turbellaria (as here) by some authorities, but the temnocephalids are treated as a separate class by others. It is now generally recognized that the monogeneans, previously in a class of the Trematoda, are sufficiently divergent from the digeneans, nearer to the cestodes and turbellarians, to be placed in a separate class (Monogenea). The classification of the digeneans used here is based on the development of the excretory bladder and associated structures. Such a scheme, however, clearly raises taxonomic problems of identification for species whose life cycles are not fully known. (J.D.Sm.)

BIBLIOGRAPHY. Classic studies of platyhelminths are LIBBIE HENRIETTA HYMAN, *The Invertebrates,* vol. 2, *Platyhelminthes and Rhynchocoela* (1951); S. YAMAGUTI, *Systema Helminthum* (1958–), taxonomic works with detailed keys—vol. 1 was revised as *Synopsis of Digenetic Trematodes of Vertebrates* (1971); and "Plathelminthes, mésozoaires, acanthocéphales, némertiens," in P.P. GRASSE (ed.), *Traité de zoologie,* vol. 4 (1961). NATHAN W. RISER and M. PATRICIA MORSE (eds.), *Biology of the Turbellaria* (1974), provides more current information. Collections of papers examining various aspects of turbellarians are found in ERNEST R. SCHOCKAERT and IAN R. BALL (eds.), *The Biology of the Turbellaria* (1981); and PETER AX, ULRICH EHLERS, and BEATE SOPOTT-EHLERS (eds.), *Free-Living and Symbiotic Plathelminthes* (1988). BEN DAWES, *The Trematoda* (1946, reissued 1968), is a classic work but somewhat out-of-date; it may be updated by two newer works, DAVID A. ERASMUS, *The Biology of Trematodes* (1972); and J.D. SMYTH and D.W. HALTON, *The Physiology of Trematodes,* 2nd ed. (1983), a detailed monograph for students. Also useful is G.R. LA RUE, "The Classification of Digenetic Trematoda: A Review and a New System," *Experimental Parasitology,* 6:306–344 (1957). Cestodes are examined in ROBERT A. WARDLE and JAMES ARCHIE MCLEOD, *The Zoology of Tapeworms* (1952, reissued 1968), a classic work, largely taxonomic, but somewhat out-of-date; J.D. SMYTH, *The Physiology of Cestodes* (1969), a detailed study for students; ROBERT A. WARDLE, JAMES ARCHIE MCLEOD, and SYDNEY RADINOVSKY, *Advances in the Zoology of Tapeworms: 1950–1970* (1974); and J.D. SMYTH and D.P. MCMANUS, *The Physiology and Biochemistry of Cestodes* (1989). Guides to identification include IAN R. BALL and T.B. REYNOLDSON, *British Planarians, Platyhelminthes, Tricladida* (1981); and GERALD D. SCHMIDT, *CRC Handbook of Tapeworm Identification* (1986).
(J.D.Sm./Ed.)

Florence

A former republic, a former seat of the duchy of Tuscany, and a former capital of Italy, Florence (Italian: Firenze; Latin: Florentia) lies almost at the centre of the Italian peninsula, some 145 miles (230 kilometres) northwest of Rome. Today it is the capital of Firenze province, covering an area of about 40 square miles (104 square kilometres). The city is surrounded by gently rolling hills that are covered with villas and farms, vineyards and orchards.

The present glory of Florence is mainly its past. Its buildings are works of art abounding in yet more works of art. The splendours of the city are stamped with the personalities of the men who made them. The geniuses of Florence were backed by men of towering wealth, and the city to this day gives testimony to their passions for religion, for art, for power, or for money. Among the most famous of the city's giants are Leonardo da Vinci, Michelangelo, Dante, Machiavelli, Galileo, and its most renowned rulers, the generations of the Medici family.

Scholars still marvel that this small city of moneylenders and cloth-makers without much political or military power rose to a position of enormous influence in Italy, Europe, and beyond. The Florentine vernacular became the Italian language; and the local coin, the florin, became a world monetary standard. Florentine artists formulated the laws of perspective; Florentine men of letters, painters, architects, and craftsmen began the period known as the Renaissance; and a Florentine navigator, Amerigo Vespucci, gave his name to two continents.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Florence was founded to control the only practicable north-south crossing of the Arno River to and from the three passes through the Apennines: one to Faenza and two to Bologna. Two thin streams, the Mugnone and the Affrico, come down through town to meet the Arno. The Affrico, not far away from its source in the Apennines, is usually a grudging gurgle amid wide gravel beds far below the quays, but sometimes it rises and swells into a powerful stream, ravaging the city with floods. The city's water supply, however, has also served as an asset, making possible the washing, fulling, and dying of cloth, resulting in the development of a major industry.

Florence's position as a major crossroads between Bologna and Rome made the city vulnerable to attack. Its hills offered some protection, but the citizens nonetheless felt compelled to erect imposing walls during the period 1285–1340; although the walls were largely torn down during urban expansion in the 1860s, their former presence remains clearly visible in a girdle of roads around the original city. Moreover, because the hillier south bank of the Arno has prevented urban growth, segments of the walls are preserved.

Beyond the historic centre of Florence, the city has expanded over the past century to accommodate waves of migration. Vast modern housing projects have been constructed, such as those at Isolotto (1954–55). These peripheral zones have actually grown to dominate the city centre, creating a kind of "open urban system"—and a vast and successful industrial district—that stretches northwest to Prato and southeast as far as Arezzo. Huge satellite towns such as Scandicci have grown to rival the centre of Florence itself.

The climate. Florence's location in a small basin encircled by hills is a determining factor for its changeable climate. Summers tend to be extremely hot and humid, and winters cool and wet. The average monthly temperature for July and August is about 73° to 75° F (23° to 24° C), with an average daytime high of about 95° F (35° C); the average monthly temperature for January is 41° F (5° C). Yet winters tend to be short-lived, ending generally in mid-March, and bring rain rather than snow. Unpleasantly cold showers can, however, persist into April, much to the discomfort of the throng of Easter tourists. The most delightful seasons in which to visit Florence are late spring and fall, when the sky becomes an azure vault and the sun warms but does not scorch.

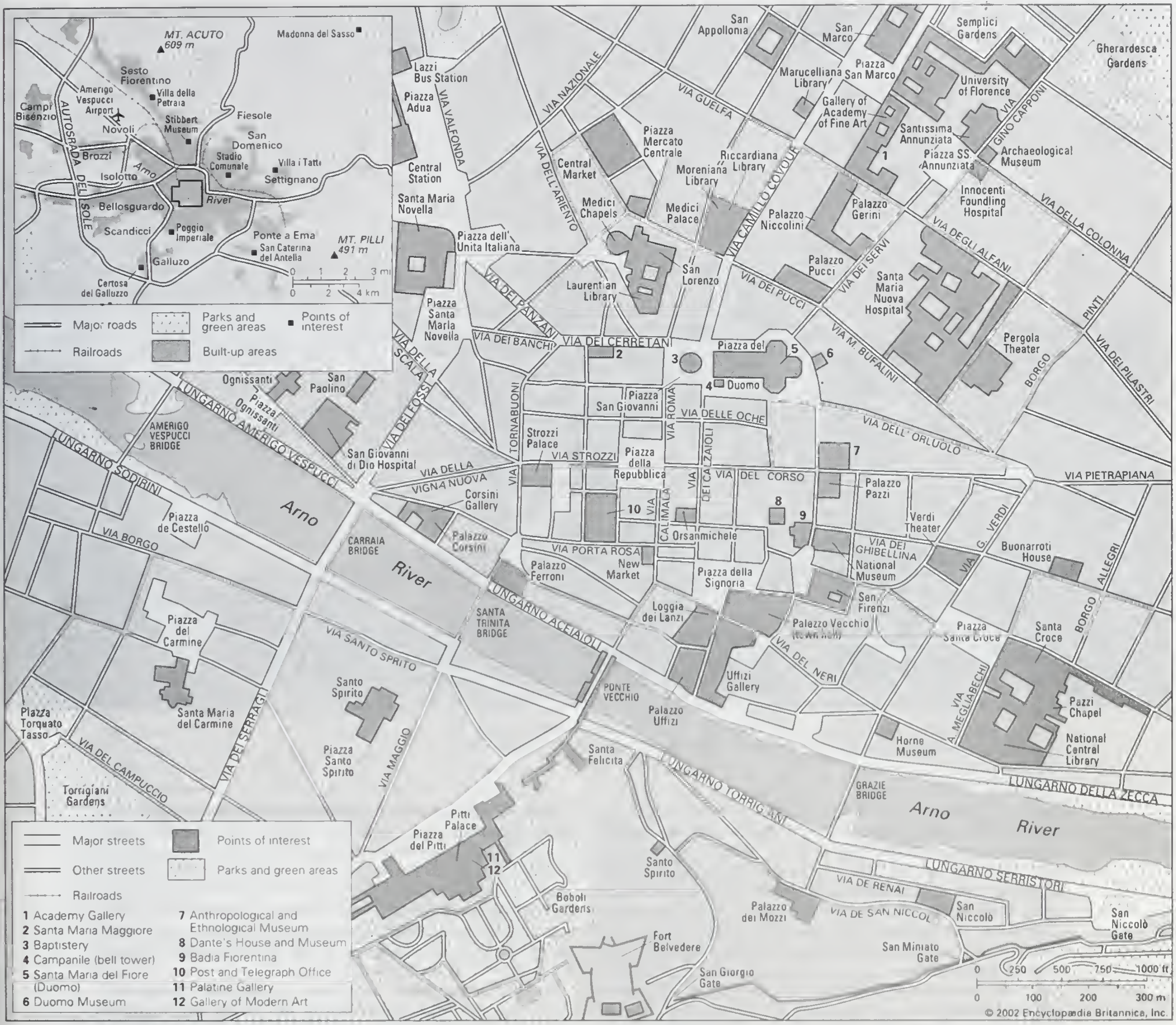
The city layout. Although most of the city of Florence was a creation of the nascent Renaissance era, the city's Roman beginnings as a typical *castrum,* or garrison town, can still be perceived. They are visible in the rectilinear grid whose axis is the Via Calimala, with a forum in today's Piazza della Repubblica (used as a market during most of its history). The skyline, however, is dominated by two imposing structures of later centuries. One of them is the austere tower of the Palazzo Vecchio (Old Palace), begun in 1299, in the Piazza della Signoria. It housed the legislative and executive branches of the local civic government (the priors) and even today functions as the town hall of Florence. Always a kind of nerve centre of local pride and power, the building was ornamented with major works of Florentine sculpture; foremost among these was Michelangelo's towering statue of "David" (today replaced by a copy). Also framing the Piazza della Signoria is the elegant Loggia dei Lanzi, built in the late 14th century; today it serves as an open-air museum for masterworks of sculpture, including Benvenuto Cellini's "Perseus."

Roman beginnings

From behind the loggia and from the flank of the palazzo, the tall, colonnaded, twin wings of a later building, the Uffizi, stretch down to the Arno. An elegant edifice designed by Giorgio Vasari, it was begun in 1560 to house the grand ducal offices. In 1574 Grand Duke Francesco I ordered the top story converted to display the Medici art treasures. The Uffizi's collection, one of the most precious in the world, offers examples of painting from the 13th century through the 18th and includes most of the significant names in Florentine art.

Duomo and campanile

The second distinctive feature of Florence's skyline is the dome of the cathedral (Duomo), Santa Maria del Fiore. The building itself, located due north of the Piazza della Signoria, was begun by the sculptor Arnolfo di Cambio in 1296. Numerous local artists continued to work on it during the following century and a half. The painter Giotto designed its sturdy bell tower (campanile) in 1334. Yet, the massive octagonal cupola (1420–36) that truly dominates



© 2002 Encyclopædia Britannica, Inc.

Florence and (inset) its metropolitan area.

both the church and the city was the proud achievement of Filippo Brunelleschi, master architect and sculptor. Opposite the cathedral stands the Baptistery; the building dates from the 11th century but was believed by Florentines to be a surviving Roman monument when they commissioned for it a series of bronze doors with relief sculptures (1330; 1401–52). The third pair of these doors, by Lorenzo Ghiberti, were of such rare beauty that Michelangelo christened them the "Gates of Paradise."

Around the perimeters of historical Florence lie the vast "newcomer" churches of the mendicant orders: to the west, Santa Maria Novella (begun 1279) of the Dominicans; to the east, Santa Croce (begun 1294) of the Franciscans. Each of these churches is a monument of Renaissance art in its decoration. The interior of Santa Maria Novella contains the Spanish Chapel, with frescoes by Andrea da Firenze; the Green Cloister, with frescoes by Paolo Uccello; the Strozzi Chapel, with frescoes by Filippino Lippi; and the Cappella Maggiore, with frescoes by Domenico Ghirlandaio, in addition to Masaccio's awe-inspiring fresco "The Trinity," with its fully realized use of perspective. The facade of Santa Maria Novella was completed (1456–70) by the design of Leon Battista Alberti. Alongside Santa Croce, Brunelleschi appended the Pazzi Chapel, designed geometrically around the motif of a circle within a square. Inside Santa Croce one finds major fresco cycles by the most famous early Florentine painter, Giotto. Paradoxically, the patrons of this church were among the richest families of Florence, despite (or perhaps because of) the vows of poverty sworn by the Franciscan order. Santa Croce has historical significance as well, because it became a kind of pantheon containing the tombs of famous Florentine scholars, writers, artists, and patriots. Across the Arno lies the modest Carmelite church of Santa Maria del Carmine, whose Brancacci Chapel displays some of the most powerful early 15th-century frescoes by Masaccio and Masolino (*c.* 1425–27).

Between the Piazza della Signoria and the cathedral lies a remarkable building, the Orsanmichele (oratory of St. Michael). In 1290 Arnolfo built a loggia here for the wheat market, which, however, was destroyed by fire; a larger loggia was erected in 1377 and then enclosed to form a church in 1380. Its chief fame comes from its early 15th-century decoration donated by the major guilds of Florence. Each guild was assigned one of the tabernacles on the exterior of the Orsanmichele and expected to commission a sculpture for it. The best works produced include bronzes of St. John the Baptist (patron of the city and of the Calimala guild) and St. Matthew (for the Cambio, or bankers) by Ghiberti and marbles of St. Mark (linen drapers) and St. George (armourers) by Donatello.

North of the cathedral lay the province of the eventual

Central Florence, seen from the southeast, across the Arno River. Rising above the city are the tower of the Palazzo Vecchio (left), the Duomo (centre), and the church of Santa Croce (right).

Graphic House, Inc

rulers of Florence, the Medici, a family of bankers. On the square behind the house of the Medici stands the Augustinian church of San Lorenzo, for which Brunelleschi made an austerely simple geometric Renaissance design based on his study of early Christian basilicas in Rome (1421). Medici patronage led to decisive artistic decorative additions. Donatello provided a bronze pulpit, and Brunelleschi added a sacristy (the Old Sacristy); about one century later Michelangelo balanced it with the New Sacristy, which contains his famous Medici Tombs. Michelangelo also designed the Laurentian Library, next to San Lorenzo, to house the great library assembled by the Medici family. Near the church sits the Medici Palace, built by the architect Michelozzo beginning in 1444. Inside, a chapel contains a fresco by Benozzo Gozzoli depicting the "Procession of the Magi" (1459), in which the followers of the Magi are given features of the Medici.

The Strozzi Palace

The grandest palace in Florence is the Strozzi Palace, begun in 1489 for one of the city's largest and wealthiest families (which, however, had been eclipsed politically by the Medici). Its enormous scale deliberately surpassed that of the Medici Palace. Noteworthy within the Strozzi Palace is a spacious courtyard, which by its use of arches and a loggia achieves a feeling of openness and simplicity.

South of the Arno lies the Pitti Palace; this grandiose structure was created for the grand duke Cosimo I by the sculptor Bartolommeo Ammannati, who extended (1558–70) a palace belonging to the Pitti family of a century earlier. The hills behind this massive palace were transformed into magnificent gardens, the Boboli Gardens, filled with fountains, statues, and an amphitheatre; here operas and concerts for the Medici rulers betokened their courtly existence as the absolute rulers of the city. Ammannati also designed a bridge to the palace, the Santa Trinità Bridge (1567–69; restored).

THE PEOPLE

Florence's greatest poet, Dante, harshly characterized his city's people as tightfisted, envious, and haughty. A touch of this severe judgment still clings to the Florentines, in whose makeup one tends to miss the exuberance and warmth associated with Italians in other towns and regions. Perhaps the Florentines, many of whom are descendants of long lines of Florentines, are reserved in self-defense against the massive stream of tourists; every year about two and a half million visitors crowd the historic sections of Florence, filling the air with foreign tongues.

The city's population increased significantly during the 20th century. Immigrants before the 1970s were mainly from the Tuscan region but also from the south of Italy. Many Chinese actually had arrived earlier, and since the mid-1970s the city and its region have attracted other people from outside Italy who found work in the area's tourist-linked service economy. These immigrants have begun to change the cultural makeup of the city. Indeed, some of the first explosions of racial animosity in Italy took place in Florence in the early 1990s, when Italian locals organized raids on immigrant street vendors, which led to a national debate over immigration.

THE ECONOMY

Industries. Thousands of Florentines work in industrial suburbs, where they are engaged in the production of furniture, rubber goods, chemicals, and food. Yet the city lives primarily from tourism and from the traditional handicrafts—glassware and ceramics, wrought iron, leatherwork, wares of precious metals, art reproductions, and the like—with some high-fashion clothing and shoe production. Key fashion companies operating in the city include Gucci and Ferragamo. Florence hosts numerous fairs throughout the year, including an international antiques fair, international fashion shows, and countless artisans' exhibits. Commercial and cultural interests blend in the city's offerings of festivals of music, opera, and the visual arts. In particular, the annual Maggio Musicale ("Musical May" festival) attracts visitors from far beyond the city. Of special appeal are the traditional festivals, many of them resplendent with the trappings of medieval pageantry and procession. Among the more famous ones are the celebrations in honour of the city's patron saint, St. John the Baptist. Visitors can watch the fireworks on June 24 (St. John's Day) or attend the "football game" staged in 16th-century costumes in the Boboli Gardens during St. John's Week.

The merchants of Florence

Craftwork is sold throughout the city, but several traditional marketplaces still exist. The vendors of straw objects—from tiny figurines to full-sized dresses—have their stalls in the Loggia of the Mercato Nuovo (New Market, built in 1547–51). Goldsmiths, silversmiths, and jewelers are concentrated on the Ponte Vecchio, one of the world's most famous bridges and the symbol of Florence. They opened for business there in the 16th century, when Grand Duke Ferdinand I deemed it inelegant for butcher shops to line the bridge as they had for the previous 200 years. He ordered practitioners of the "vile arts" to give way to workers in precious metals. The new occupants eventually enlarged their shops by building outward over the water, propping their three-story additions on brackets from the bridge. The back elevations of these extensions give the bridge its picturesque air. Above the shops a covered passage was constructed in 1564–65 to connect Cosimo's palace (the Pitti Palace) on the left bank with the newly erected government offices (the Uffizi) on the right bank.

Artisans who fashion the gold, silver, jewelry, straw, intarsia (inlaid woodwork designs), leather goods, glass, pottery, and embroidery complain of being squeezed out of existence by the pressures of modern economic life. These artisans can, however, still be seen through the open doors of their workrooms, engaged in the tasks and poised in the attitudes shown in the carvings on the 15th-century facade of the guildsmen's church, Orsanmichele.

Traditional heavy industry is still important in the area. Major employers include Nuovo Pignone (now part of the American General Electric Company), maker of steam turbines and compressors, and Piaggio & Company (located in and around Pisa, 50 miles to the west), maker of the famous Vespa scooter. The city is now part of a huge industrial district running northwest to Prato and Pistoia. This zone, with its small businesses and quality export production, was one of the centres of the prosperous "third Italy" of the 1990s, rivaling similar zones in Emilia-Romagna and Veneto in employment and profits. Hundreds of thousands of former sharecroppers from rural areas of Tuscany have become small businessmen in a single generation, avoiding the trauma of "normal" rapid industrialization. However, the environment has suffered as the beautiful Tuscan countryside has slowly been urbanized and as motor-vehicle traffic has threatened to suffocate not just the city but the entire region.

Transportation. In the central area of Florence a solid pair of walking shoes is the best mode of transportation, especially since the historic section has been closed to motor vehicles. Buses and taxis are also available, as are bicycles for hire. The main highway, the Autostrada del Sole, passes west and south of the city. Because Florence lies on the country's main north-south train line, rail connections are highly dependable and efficient. The Eurostar connects Florence with Milan in less than three hours and with Rome in less than two. In addition, Florence has its own airport, Amerigo Vespucci (formerly Peretola), only three miles from the city centre. It is too small for intercontinental traffic, but Pisa's Galileo Galilei International Airport is an hour's train journey away.

CULTURAL LIFE

Florence has numerous museums, mostly devoted to painting and sculpture. The National Central Library (Biblioteca Nazionale Centrale) has been the Italian library of deposit since 1870, receiving a copy of every book published in the nation. It houses millions of autographs, manuscripts, letters, incunabula, and books, including many rare editions. The Riccardiana and Moreniana libraries adjoining the Palazzo Medici-Riccardi have the most complete collection, including valuable manuscripts, of works on Tuscan history. The Gabinetto Scientifico e Letterario G.B. Vieusseux is a scientific and literary library founded in 1819 by Jean-Baptiste Vieusseux, who was the central figure of a group that included the leading literary figures of Italy at that time.

Universities and learned institutions

After Lorenzo de' Medici transferred the University of Florence (founded 1349) to Pisa in 1472, the medical school remained behind, leading the scientific movement in Italy and forming the nucleus for the university that was legally constituted only in 1923. The Academy of the Crusca was established in 1582 to prepare an Italian dictionary; *crusca* means "bran," the academy's symbol is a sieve, and its object remains to winnow impurities from the language. Other specialized learned institutions include an observatory; academies of fine arts, science, letters, and agrarian economics; and institutes of Etruscan and Italian studies, of the history of art, and of the history of optics. The Italian Dante Society, the Italian Botanical Society, and the Society for Geographical Studies are in Florence.

An increasing number of foreign nations and universities conduct study institutes in Florence. The Harvard University Center for Italian Renaissance Studies is located at the exquisite Villa i Tatti, bequeathed by the art historian Bernard Berenson, in the hills at Settignano. The Universities of Grenoble (Fr.) and Paris, Syracuse University (N.Y., U.S.), Stanford University (Palo Alto, Calif., U.S.), Smith College (Northampton, Mass., U.S.), universities of The Netherlands, and the U.S. state universities of California are represented.

The glory of many Florentines is the city's football team, Fiorentina—or "la Viola," as the team is affectionately called, owing to the players' purple shirts. The club has won the Italian championship on only two occasions (in 1956 and 1969), but it continues to inspire fanatic support from its followers. When star Roberto Baggio was sold to archrival Juventus of Turin in 1990, Fiorentina supporters caused riots that paralyzed the city. The stadium, originally designed in the 1930s by modernist architect Pier Luigi Nervi, has achieved national monument status. It was refurbished for the 1990 World Cup and renamed the Stadio Comunale "Artemio Franchi," or simply Franchi Stadium.

History

THE EARLY PERIOD

Founding and growth. Florentia ("The Flourishing Town") was founded in 59 BC as a colony for soldiers of the armies of Rome and was laid out as a rectangular garrison town (*castrum*) below the hilltop Etruscan town of Faesulae. Its streets formed a pattern of rectangular blocks, with a central forum, a temple to Mars, an amphitheatre, and public baths. By the 3rd century AD Florence was a provincial capital of the Roman Empire and a prosperous commercial centre. During the early medieval centuries, Florence was occupied chiefly by outsiders: first by Ostrogoths in the 5th century, then by Byzantines in the 6th century, and eventually by Langobards, or Lombards. From the late 10th century onward Florence prospered, and, under the rule of Countess Matilda of Tuscany (1069–1115), it became the leading city in Tuscany.

In 1293 Florence adopted a constitution called the Ordinances of Justice, which barred both the nobility and labourers from political power. It also provided for frequent changes of office to ensure that no group or individual could get control of the state; thus the nine priors who constituted the Signoria (the governmental body) were each elected for a mere two months. As a result, Florentines developed a keen interest in their politics and became a community of civil servants available for public life; but the lack of continuity often provoked factional intrigues and alliances.

Growing economic and political power

During the 12th and 13th centuries the economic and political power of the city grew steadily. The rise of the Florentine woolen cloth industry and of banking provided a basis of capital. Then the resolution in 1266 of a bitter strife between two internal factions oriented respectively toward papal (Guelf) and imperial (Ghibelline) protection resulted in victory for a group of Guelf merchant families in the city (as well as the exile in 1302 of Florence's greatest poet, Dante Alighieri). They took over papal banking monopolies from rivals in nearby Siena and became tax collectors for the pope throughout Europe. From such a foundation, Florentine families, led by the Bardi and the Peruzzi, came to dominate both banking and international merchant business. Locally, Florence also added neighbouring cities to its sphere of influence and obliged rival powers—Pisa, Siena, Pistoia, and Arezzo—to become its allies.

With a balance between its leading merchant families, Florence was now ruled by its guilds, divided into seven major guilds and a number of minor ones. The city's *podesta,* or chief magistrate and police chief, could be selected only from the major guilds. Political parties grew up along the issues of aggressive expansion and preservation of peace; the former policy was embraced by the Blacks (Neri; the rich merchants), the latter by the Whites (Bianchi; the lesser citizens).

Just before the middle of the 14th century, Florence had become a metropolis of about 90,000 people, making it one of the great cities of Europe (alongside Paris, Venice, Milan, and Naples). However, in the summer of 1348 the Black Death struck, reducing the population by half. The city's ordeal during this period has been vividly portrayed by the chronicler Matteo Villani and by the writer Giovanni Boccaccio in the preface to his stories of the *Decameron.* The bankruptcies of the Bardi and the Peruzzi a few years before the Black Death had already shaken the city's prosperity, and it never fully recovered from these double disasters. Famine and renewed bouts of the plague continued throughout the 14th century, sparking unrest among the politically unrepresented population. In 1378 a proletarian rebellion of the cloth workers, the Ciompi revolt, was put down by an alliance of merchants, manufacturers, and artisans. The economy of the city remained depressed, and the rivalry of adjoining polities, first Milan and then Naples, only intensified the threats to Florence's prosperity in the early 15th century. One of the few successes was the conquest of Pisa in 1406, making Florence a maritime power at last. Partly in self-defense, Florence became a major territorial power alongside Venice, Milan, and Naples.

During this period of adversity, the power of the guilds and their domination of the city were on the wane; as a result, successful merchants and bankers, chiefly Cosimo de' Medici and Giovanni Rucellai in the 15th century, were able to shape civic politics and culture through a system of oligarchy and patronage. They underwrote the accomplishments that are now singled out with the term "Renaissance," and their palaces came to dominate the city as fully as the church buildings in which they established their family chapels.

The rule of the Medici. Cosimo de' Medici (Cosimo the Elder; d. 1464) became the leading citizen in Florence after his return in 1434 from a year of exile. He achieved this position by virtue of his great wealth (the result of the largest banking network in Europe) and an extensive network of patronage obligations. While he never accepted public office, his faction dominated the city. He lived an increasingly opulent life, as is apparent in the ostentation of the Medici Palace and the patronage of churches such as San Lorenzo and the Monastery of St. Mark, with its frescoes by Fra Angelico. Investment in culture, including the patronage of artists and architects and the purchase of books and manuscripts, became a fundamental expression of the Medici's aristocratic way of life; it was continued by Cosimo's son, Piero, and his grandson, Lorenzo (d. 1492; dubbed "the Magnificent"). In all but name, Florence was now ruled by a Medici prince, whose position resembled that of the tyrants in other Italian cities such as Milan, Ferrara, Mantua, and Urbino.

Girolamo Savonarola

Stability was briefly threatened in 1478 by the brutal but abortive Pazzi conspiracy seeking to end the Medici rule. In 1494, shortly after the death of Lorenzo, French armies under King Charles VIII invaded Italy. They were backed against the Medici by the popular party in Florence, which (with French help) succeeded in exiling the Medici and declaring Florence a republic. The consequence, however, was the loss of political autonomy to the larger conflicts of Italian peninsular struggles. Republican Florence was led briefly by a fiery Dominican preacher, Girolamo Savonarola, who boldly condemned the luxury and urbane culture of his predecessors. His strict rule came to an end in 1498, but with it closed a phase of Florentine greatness.

The Medici returned to Florence in triumph in 1512 behind the papal and Spanish armies, reasserting power in a clear and ruthless manner. (Such an unambiguous pursuit of power by leaders at this time was given codification in 1513 by Niccolò Machiavelli in his treatise *The Prince.*) In addition, the younger son of Lorenzo was elected Pope Leo X; his pontificate (1513–21) was noteworthy for its cultivation of the arts, especially his employment of Raphael. Leo was shortly followed by another Medici pope, Clement VII (1523–34). However, in 1527 the riotous Spanish army of Emperor Charles V overran Rome, and during this moment of weakness republicans again expelled the Medici from Florence, only to be punished in 1530, when pope and emperor were reconciled. Then in 1536 the statesman and historian Francesco Guicciardini began to compose his *History of Italy,* with its ideal vision of the era of Lorenzo the Magnificent and its pessimism concerning more recent events. In 1537 Charles V installed Cosimo de' Medici (Cosimo I; d. 1574) as official duke of Florence (grand duke of Tuscany after 1569). Cosimo and his wife, Eleonora of Toledo, patronized the arts and undertook vast building programs, such as the construction of the Uffizi, the renovation of the Palazzo Vecchio, and the reconstruction of the Pitti Palace.

With the rise of Cosimo I to titled nobility and to absolute rule in Florence, the political and cultural vitality of the city had all but ebbed, prompting a modern scholar to refer to the succeeding era as the "forgotten centuries." Florence's dukes had become minor players in the broader European balance of great powers, and they linked themselves chiefly with the noble houses of France. Marital alliances of Medici family members with members of the French nobility include Catherine de Médicis (d. 1589), queen of Henry II and later regent of France; Grand Duke Ferdinand I (d. 1609), who married Christine of Lorraine; and Marie de Médicis (d. 1642), who married King Henry IV of France. The city generally declined under prolonged Medici rule, a process that was marked only by the extended reign of Cosimo III (1670–1723) and the end of the family with the death of his son, Gian Gastone (d. 1737).

From outside rule to unification. After the rule of the Medici, Florence was governed from outside, as Francis Stephen of Lorraine, the husband of Empress Maria Theresa of Austria, became the grand duke of Tuscany. Following a Napoleonic interlude, Leopold II of Habsburg was the last outside ruler (1824–59). He eventually abdicated in favour of the new Italian king, Victor Emmanuel. In 1860 Florence annexed itself to the new kingdom of Italy, serving as its capital during the period 1865–70.

From the late 18th to the mid-20th century, a large Anglo-American colony was an integral part of the Florentine scene. The poet Elizabeth Barrett Browning, who is buried in the small English cemetery, noted that the city was "cheap, tranquil, cheerful and beautiful." The Horne Museum, near Santa Croce, and the Stibbert Museum, in the north, are examples of houses and collections left by foreigners to their adopted city.

EVOLUTION OF THE MODERN CITY

Human and natur despoliation

During the decades after unification, much of Florence's past was seriously jeopardized as a huge debate surrounded the renewal of the city. Many of its finest structures were altered or defaced, its medieval walls were largely pulled down, and its ancient centre was laid waste as entire zones were cleared and replaced with squares and other public spaces. Florence's brief period as Italy's capital in 1865–70 saw a vast increase in the city's population. Growth continued chaotically into the 20th century, when order and sense were slowly restored to the renovations.

Florence was occupied by the German army in the latter stages of World War II. Antifascist resistance was very strong in the city and Tuscan region, and fighting was heavy. On leaving Florence, the Germans blew up all the bridges across the Arno but spared the famous Ponte Vecchio. After the war, debates again arose regarding the reconstruction of the city. The 1950s saw Florence expand into its periphery, and in 1962 a plan was developed partly to guide expansion away from the Florence-Prato conurbation. Despite certain innovative qualities, the plan failed, and the two cities are now all but one, constituting one big urban sprawl. The population of Florence reached more than 400,000 in the 1980s but soon fell slightly below that level.

In addition to human despoliation, nature has been an occasional adversary of Florentine life, mainly in the form of floods of the Arno. The city's bridges were destroyed in the 12th century and again in 1333. In 1557 the Ponte Vecchio held fast, but the others were destroyed. The most devastating occurrence was the flood of November 1966. The city's cultural heritage was grievously damaged by waters coursing through the streets and swirling into buildings, depositing debris, mud, and oil. Major public works took place to try to ensure that there would be no repetition of this tragic event. Despite some scares, the Arno has not burst its banks since then.

Since the mid-20th century, Florence has been governed largely by centre-left administrations that have been known for their reformist zeal, honesty, and efficiency (with some notable exceptions). Through the popular mayor Giorgio La Pira, left-wing Catholics were able to run the city for a number of years in the late 1950s and early 1960s. Since then the subcultures of the left have survived in Florence, despite their erosion elsewhere in Italy, and the old Italian Communist Party, now known as the Partito Democratico di Sinistra (Democratic Party of the Left), has managed to maintain a strong presence as well. With its large university population and radical-left traditions, Florence was one of the centres of student and worker revolts in 1968.

In the 1980s and 1990s, the main problems of the city were environmental. The huge influx of tourists threatened to swamp the city in every sense, and the city centre began to lose its distinctive character and its artisan workshops. Pollution reached record levels, thanks to mass ownership of motor vehicles and to the peculiar geographic position of the city. Drastic measures were finally taken in the 1980s with the closure of the city centre to private cars. This policy has gone some way toward alleviating pressure on the city, but the area is still suffocated by thousands of scooters and buses, and the city still endures a constant tourist onslaught. Florence faces the danger of becoming a commercialized provincial town, full of trinket shops, overpriced bars, and crowded museums. Certainly the dynamic economic district running across to Prato is not matched by the city itself, which continues to capitalize on the past. The label of "museum-city" is thus perhaps best applied to the Florence of today, despite the highly productive region of which it is the capital.

BIBLIOGRAPHY

General works. Numerous guides to the city are published every year, most of them repeating basic information. A useful example is PAOLO DE SIMONIS *et al.*, *Florence: A Complete Guide to the Renaissance City, the Surrounding Countryside, and the Chianti Region,* trans. from Italian by ANTONY SHUGAAR (1999), in the *Heritage Guide* series. An excellent guide to architecture in the city can be found in *Architectural Guides: Florence,* 2nd ed. (1998), in the *Allemandi's Architectural Guides* series. A more scholarly work is GUIDO ZUCCONI, *Florence: An Architectural Guide* (1995). ALTA MACADAM, *Florence,* 8th ed. (2001), in the *Blue Guide* series, provides a general introduction to Florence and its history as well as maps and tours of different sections of the city. MARY McCARTHY, *The Stones of Florence* (1959, reissued 1987), offers a more evocative and literary survey. A useful Italian guide is ADRIANO AGNATI (ed.), *Firenze: le colline, il Mugello, il Valdarno, il Chianti* (1994), published by Touring Club Italiano.

History. Not surprisingly, most histories of Florence focus on the origins and developments of the Renaissance era, particularly the 14th and 15th centuries. The early era of civic greatness is explored in GEORGE HOLMES, *Florence, Rome, and the Origins of the Renaissance* (1986, reissued 1989). The city received its own 16th-century history at the hands of one of its greatest political thinkers, NICCOLÒ MACHIAVELLI, *History of Florence and of the Affairs of Italy: From the Earliest Times to the Death of Lorenzo the Magnificent* (1994; originally published in Italian, 1532), available in many printings. Shortly afterward this account was complemented by the authority of FRANCESCO GUICCIARDINI, *The History of Florence* (1970; trans. from Italian, 1931; originally published in 1859). The historiography of both chroniclers is the subject of FELIX GILBERT, *Machiavelli and Guicciardini: Politics and History in Sixteenth-Century Florence* (1965, reprinted 1984); supplementing that is GISELA BOCK, QUENTIN SKINNER, and MAURIZIO VIROLI (eds.), *Machiavelli and Republicanism* (1990, reissued 1993). An introduction to Machiavelli is available in QUENTIN SKINNER, *Machiavelli: A Very Short Introduction,* rev. ed. (2000).

An impressive analysis of social and economic conditions, as well as of the political situation, is presented in GENE BRUCKER, *Florentine Politics and Society, 1343–1378* (1962), *Renaissance Florence* (1969, reprinted 1983), and *The Civic World of Early Renaissance Florence* (1977); and GENE BRUCKER (ed.), *The Society of Renaissance Florence: A Documentary Study* (1971, reissued 1998). Issues of daily life (and death) are discussed in C.C. BAYLEY, *War and Society in Renaissance Florence: The De Militia of Leonardo Bruni* (1961); KATHARINE PARK, *Doctors and Medicine in Early Renaissance Florence* (1985); and DANIEL R. LESNICK, *Preaching in Medieval Florence: The Social World of Franciscan and Dominican Spirituality* (1989). More on economic and social conditions is in RICHARD A. GOLDTHWAITE, *The Building of Renaissance Florence: An Economic and Social History* (1980, reissued 1990); and CHRISTIANE KLAPISCH-ZUBER, *Women, Family, and Ritual in Renaissance Italy,* trans. from French (1985, reissued 1987), the latter emphasizing Florence's female population. Developments in public culture are observed in RICHARD C. TREXLER, *Public Life in Renaissance Florence* (1980, reissued 1991). Neglected post-Renaissance developments are discussed in the aptly titled ERIC COCHRANE, *Florence in the Forgotten Centuries, 1527–1800: A History of Florence and Florentines in the Age of the Grand Dukes* (1973). Other aspects are explored in JEAN-CLAUDE WAQUET, *Corruption: Ethics and Power in Florence, 1600–1770* (1991; originally published in French, 1984).

A more recent history by SIDNEY TARROW, *Democracy and Disorder: Protest and Politics in Italy, 1965–1975* (1989), covers events of 1968, in particular at Isolotto. Italian books on the development of Florence in the 20th century are LAURA CERASI, *Gli ateniesi d'Italia: associazioni di cultura a Firenze nel primo Novecento* (2000); CARLO CRESTI, *Firenze, capitale mancata: architettura e città dal piano Poggi a oggi* (1995); GIORGIO MORI (ed.), *La Toscana* (1986); and GIORGIO SPINI and ANTONIO CASALI, *Firenze* (1986).

Art. Florentine artists are the chief heroes in the later 16th-century history of Italian art by GIORGIO VASARI, *Lives of the Most Eminent Painters, Sculptors & Architects,* trans. by GASTON DU C. DE VERE, 10 vol. (1912–15, reprinted 1976; originally published in Italian, 3 vol., 1550); and much detail on the same period emerges from an artistic autobiography by BENVENUTO CELLINI, *The Life of Benvenuto Cellini,* trans. by JOHN ADDINGTON SYMONDS (1995; originally published in Italian, 1728), available in many printings. Many general histories of Italian Renaissance art devote considerable space to Florentine developments, with some giving special emphasis to the city alone. Outstanding among these is MARTIN WACKEMAGEL, *The World of the Florentine Renaissance Artist: Projects and Patrons, Workshop and Art Market,* trans. by ALISON LUCHS (1938, reissued 1981; originally published in German, 1938); it can be supplemented by the more argumentative FREDERICK ANTAL, *Florentine Painting and Its Social Background: The Bourgeois Republic Before Cosimo de' Medici's Advent to Power, XIV and Early XV Centuries* (1948, reprinted 1986); and the more general study of the arts in PETER BURKE, *The Italian Renaissance: Culture and Society in Italy,* 2nd ed., rev. (1999). The chief pictorial handbook remains BERNARD BERENSON, *Italian Pictures of the Renaissance: A List of the Principal Artists and Their Works, with an Index of Places: Florentine School,* 2 vol. (1963). Frescoes are the subject of EVE BORSOOK, *The Mural Painters of Tuscany: From Cimabue to Andrea del Sarto,* 2nd ed., rev. and enlarged (1980). WALTER PAATZ and ELISABETH PAATZ, *Die Kirchen von Florenz,* 6 vol. (1952–55), is an informative work on churches. Sculpture is covered in JOHN POPE-HENNESSY, *An Introduction to Italian Sculpture,* 4th ed., 3 vol. (1996, reissued 2000); and CHARLES SEYMOUR, *Sculpture in Italy: 1400–1500* (1966). An appreciation of a high point of Florentine painting is offered in S.J. FREEDBERG, *Painting of the High Renaissance in Rome and Florence,* new rev. ed., 2 vol. (1985); and ANDRÉ CHASTEL, *The Flowering of the Italian Renaissance,* trans. from French (1965). Art in its intellectual or phenomenological context is discussed in MICHAEL BAXANDALL, *Giotto and the Orators: Humanist Observers of Painting in Italy and the Discovery of Pictorial Composition, 1350–1450* (1971, reprinted 1986), and *Painting and Experience in Fifteenth-Century Italy: A Primer in the Social History of Pictorial Style,* 2nd ed. (1988). (B.E./L.A.Si./J.Ft.)

Folk Arts

Applications of the term folk have become problematic. On the face of it, the word has a simple definition—*i.e.,* "the common people, especially those of rural areas." But, like all class distinctions, "folk" carries with it the complications of historical context and point of view. Such a designation formerly implied a high degree of cultural insularity, but mass culture has since penetrated virtually every social and geographic area of the industrial world. Romantic and often patronizing views of folk life have further distorted popular perceptions. A relative and subjective concept like "folk," while academically useful, resists a consensus of interpretation. Anthropologists and folklorists have classified as "folk art" a broad range of materials, from the arts of the so-called primitive or nonliterate peoples to those of the nonelite within literate cultures, the latter variously known as folk, visionary, outsider, or naive. The folk art rubric has also been extended to include all manner of traditional or nostalgic artistic productions, even the self-consciously quaint.

For the purposes of this article, the folk arts may be considered to encompass the traditional, typically anonymous arts produced by members of a nonruling, relatively nonaffluent, often rural and uneducated stratum of industrial society. As expressions of community life, the folk arts are distinguished from the academic or self-conscious or cosmopolitan expressions that constitute the fine arts and decorative arts of the elite. As nonrepresentative of their respective societies as a whole, they are distinguished from the arts of nonliterate cultures. They are also to be differentiated from the so-called popular arts, which appeal to a mass audience and typically depend upon the mass media for their dissemination.

While in popular usage the term folklore refers almost exclusively to a single aspect of folk art, the oral literary tradition, scholars use it to embrace all of the artistic genres of folk culture. In the context of this article the terms folklore and folk art are interchangeable. The first part of this discussion treats folklore as an academic field, viewed from humanistic, anthropological, and psychological perspectives. The functions of folklore and the role of the folk artist, as interpreted from these viewpoints, are also examined. The remainder of the article considers the origins, formal characteristics, and distribution of the various genres of folk art: oral literature, music, dance, and the products of material culture (*i.e.,* the visual arts). (Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 613, 621, 622, and 624–629, and the *Index.*

The article is divided into the following sections:

Folklore as an academic discipline

No field of learning is perhaps more misunderstood than folklore. The public, and many academics, do not know that folklore is an intellectual subject with its own substantial, worldwide body of scholarship. Part of the confusion lies in the use of the word folklore to signify both the content and the study of traditional materials. Further misunderstanding results from the varying senses of folklore in different countries. In much of South America and in some European nations, folklore generally applies to public performances of song, dance, and festival, and in scholarly usage the term refers to the study of peasant culture. In the United States, the word folklore often conjures up an image of folksingers or old-timers spinning yarns of Paul Bunyan and Johnny Appleseed, who are largely contrived. The German term *Folklorismus* and the English term *fakelore* have been coined to distinguish between the genuine and the synthetic.

THREE APPROACHES TO FOLKLORE

Serious students of folklore by no means agree on the boundaries of their discipline but they tend to follow one of three prevailing perspectives.

Emphasis on narrator

Humanistic perspective. The humanistic folklorist sees the materials of folklore as "oral literature" and the folk as its artistic performers. He emphasizes the creative role of the narrator or bard, seeks information on his biography and personality, closely observes his interaction with the audience, and analyzes his total repertoire much as a literary critic assesses the achievement of the novelist, poet, or dramatist. Usually the older exponents of this approach have entered folklore from departments of modern language and literature or music or classics. A case in point is Albert Lord, professor of Slavic literature at Harvard University, whose folklore interests developed from a concern with the south Slavic oral epic, still sung by bards in Yugoslavia. Lord came to the conclusion that each epic singer constructed his narrative song by improvising around a stock of fixed images, epithets, and conventional expressions, which he alone kept firmly in mind. This "oral-formulaic" theory Lord then applied to the Homeric epics, which are now known only in their written forms but which he conjectured had been orally composed in the same manner as the contemporary Slavic epic songs. The interest in the human carrier of tradition, in his worldview and belief system, his cultural inheritance and acculturative experiences, distinguishes this species of folklorist.

Emphasis on cultural norms and values

Anthropological perspective. The anthropological folklorist examines the materials of folklore using the hypotheses of the social sciences. He looks for cultural norms and values and predictable laws of behaviour that form a consistent pattern in the nonliterate society he has closely observed. Folklore to him is an aesthetic product of this society, mirroring its values and offering a projective screen that illuminates its fantasies. Hence the folkloristic anthropologist frequently reports a one-to-one correlation between the value system and the tale repertoire of a given culture, whereas the humanistic folklorist promptly points out that the same tales are found in many parts of the globe and so can scarcely be said to reflect the ethos of a particular people, even when they have been strongly localized.

-At different periods folklore and anthropology have enjoyed an intimate relation in England and America. The father of anthropology in England, E.B. Tylor (1832–1917), drew heavily upon the materials of folklore in his two great works, *Primitive Culture* (1871) and *Researches into the Early History of Mankind* (1865), which in turn contributed to the growth of a school of so-called anthropological folklorists. The leader of this school, Andrew Lang (1844–1912), a versatile man of letters, in many clever essays and books elaborated a theory of "survivals" based on Tylor's hypothesis that from the beliefs and customs held by peasants and contemporary savages the folklorist could reconstruct the ideas of prehistoric man. In the United States, Franz Boas (1858–1942), the father of American anthropology, influenced many doctoral students, later eminent in their own right, to pay attention to folklore in their fieldwork. Boas himself considered that tribal traditions preserved an ethnological record of older culture traits and should be consulted in lieu of written documents. His disciple and successor as editor of the *Journal of American Folklore,* Ruth Benedict (1887–1948), pointed out that the tribal mythology depicted violations of taboos, such as the hero-trickster sleeping with his mother-in-law, which would never be tolerated in real life. Another student of Boas, Melville J. Herskovits (1895–1963), broke with the Boasian concentration on North American tribes to explore African cultures but retained the same emphasis on folklore and inculcated this emphasis in his own students.

American folklore-minded anthropologists all experienced difficulty in employing the term folklore within a culture almost wholly oral and traditional and resorted to various substitutes. William Bascom, a student of Herskovits, suggested the term verbal art to denote the oral aesthetic traditions of tale, proverb, song, and riddle in the culture, leaving out the supernatural-belief system and the plastic arts, which a humanistic folklorist includes in his concept. Bascom also clarified the functional uses of folklore in nonliterate societies, in accord with the anthropologists' stress on the social mechanisms that enable a society to perform its business. His book, *Ifa Divination* (1969), demonstrates how Yoruba diviners have recourse to the tribal repertoire of traditional narratives in arriving at their analyses of individual problems. The beans they throw on the Ifa board fall in a series of complex patterns that the diviners key to folktales, whose contents are then applied to the particular situation in hand. English and American anthropologists of the 1960s have in the main moved away from the soft, folkloric parts of culture to the hard, sociological data of kinship organization and political structures.

Emphasis on behaviour

Psychological–psychoanalytic perspective. The psychological–psychoanalytic folklorist views the materials of folklore neither aesthetically nor functionally but behaviouristically. Myths, dreams, jokes, and fairy tales are taken to express hidden layers of unconscious wishes and fears. Sigmund Freud drew extensively upon folk sources in such works as *The Interpretation of Dreams* (1899), *Jokes and Their Relation to the Unconscious* (1905), and *Totem and Taboo* (1913). In his view, and that of many writers on folklore influenced by him, folklore texts are to be interpreted symbolically in terms of sexual imagery and the Oedipus complex. In "Jack and the Beanstalk," the stalk is construed as a phallic symbol and Jack's chopping it down signifies a masturbation fantasy. "Little Red Riding Hood" tells a tale of a young virgin, identified by the red cap, a menstrual symbol, who wanders from the straight and narrow path, to be devoured (seduced) by the wolf in disguise as the grandmother, an Oedipal figure. In the modern college legend "The Hook," the torment suffered by a coed is multilayered: she narrowly averts assault from an escaped lunatic, one of whose arms has been replaced with a metal hook; the psychological interpretation explains the hook as the phallus and the episode as the coed's fear of the sexual act. When C.G. Jung broke with his mentor Freud and substituted the symbolism of social unconsciousness for the symbolism of sexual drives, he still retained his deep interest in myths, dreams, and tales as psychoanalytic media, and the Jung Institute in Zürich continues to offer courses on the uses of folklore in psychology.

Such a work as *The Trickster,* edited by the American anthropologist Paul Radin, with commentaries by the mythologist Károly Kerényi and by Jung, offers within a single volume three psychoanalytical interpretations of the Winnebago cycle of trickster tales. Kerényi, although he was a collaborator of Jung, takes a Freudian position and sees the trickster as a phallic figure, the spirit of disorder. Jung finds unconsciousness as the chief and alarming characteristic of the trickster, whom he considers the forerunner of the Saviour. Radin synthesizes both Freudian and Jungian analyses in his conception of the trickster as evolving from a psychically unaware individual to a socially developed being, reinterpreted by each new generation as god, hero, or buffoon. The most committed Freudian folklorist of the 1960s was Gershon Legman, who was Hungarian by birth but resided in the United States and France. In *The Rationale of the Dirty Joke* (1968), Legman properly gives attention to the most current folktale genre of modern society but looks beneath the seemingly obvious motivations of dirty jokes to violate sexual and anal taboos and imputes to them such latent impulses as male castration fears, female revulsion at the sexual act, and homosexual drives.

The three perspectives of folklore are not mutually exclusive. Anthropologists may at times employ an essentially humanistic approach, as Daniel J. Crowley has done in his detailed ethnography of the styles and repertoires of Bahamian narrators, *I Could Talk Old-Story Good* (1966). They also have shown sympathy for the psychological–psychoanalytical node. A case in point is the study of *Water-Witching, U.S.A.* (1959), by anthropologist Evon Vogt in collaboration with the psychologist Ray Hyman, who explained the widespread phenomenon whereby diviners, called dowsers, located underground water with a forked branch as an anxiety-releasing mechanism for depressed farmers. Younger folklorists of the 1960s trained as humanists have shown an increasing orientation toward social-science methods of model building and statistical techniques.

THE ORIGINS OF FOLKLORE

Grimms' Fairy Tales and the collection of folklore

When scholars and other intellectuals began to recognize early in the 19th century the presence in their midst of a vast floating body of folk traditions and practices, they promptly began to speculate on its origins. When in 1812–14 the German philologist Jacob Grimm (1785–1863), together with his brother Wilhelm (1786–1859), published the first volume of the *Kinder und Hausmärchen* (customarily translated as *Grimms' Fairy Tales*), thus initiating the science of folklore, he connected his collecting of village tales to an elaborate mythological system of origins, outlined in his *Deutsche Mythologie.* This influential work was first issued in 1835, and its fourth edition was translated into English (3 volumes, 1883–88) under the title *Teutonic Mythology.* A German nationalist, Grimm postulated a highly developed religious pantheon of pre-Roman times suppressed by the medieval church and surviving only in broken fragments of peasant beliefs and stories. The *Märchen,* or tales, were the detritus of the old myths.

In the mid-19th century, following the discovery of Sanskrit as the ancient classical language of India and the

"Disease of language" theory

parent of European tongues, a pan-Aryan theory of origins developed. In 1856 Max Müller, a German-born scholar who went to London to translate from the Sanskrit the *Sacred Books of the East* and stayed on at Oxford University, published a long essay on "Comparative Mythology," introducing the new theory. Through a process he called "disease of language," Müller conjectured that myths arose from forgetfulness of the original names of deities, transferring to them metaphorical qualities suggested by the names. Thus Dyâus, the sky deity of old India, would be thought of and narrated about as the sky, or by association, heavens, clouds, storms, and winds. The Greek myths sprang from the Indic—Zeus was the linguistic counterpart of Dyâus—and could be explained on the same grounds. Although Müller applied his theory only to the advanced Indo-European Aryan civilizations, he did point to similar myth-making among primitive peoples. Some of his followers, notably George W. Cox, carried the school of solar mythology to extreme limits and explained all folk narratives, epics, and ballads as originating in early man's poetic rendering of the conflict between the sun and the night.

Tylor added a new conception of the origins of folklore without challenging Müller's solar mythology but by pushing the starting point back beyond the Aryans to prehistoric man. Looking at the universe animistically, primitive man endowed the elements, the animals, the plants, and the rocks with personalities and souls. Such beliefs survived into modern civilization among the unlettered lower classes and formed a crust of folklore. The strongest challenge to Tylor's evolutionary theory of culture and "devolutionary" theory of folklore (in the phrase of Alan Dundes) first emerged in a hypothesis advanced by the German scholar Theodor Benfey in his introduction to the Indian story collection *Panchatantra* (1859). Benfey claimed that India, the seat of an ancient high civilization that had spread to Europe, was the home of the master tales subsequently found in the Grimms' collection. Along with language and mythology, these wonder tales had diffused from India to Europe in ancient times and again in historic times along well-traveled trade routes. Benfey's arguments persuaded other folklorists, notably Emmanuel Cosquin in France and William Alexander Clouston in Scotland, who added to his evidence of story migration from India eastward. In the past century, the primacy of ancient India as the fountain of world folklore has gradually been whittled away by the claims of other dispersal points, such as Egypt and Greece, and by the growing realization that no sweeping generalization could account for folklore origins. According to the Finnish folklorists, whose historical-geographical method attracted most scholars in the first half of the 20th century, the life history of each complex tale and ballad requires separate investigation. After exhaustively comparing the traits of all the assembled variants of a given tale type, the Finns (specifically Kaarle and Julius Krohn and Antti Aarne) believed they could establish its original form and approximate place and period of genesis. The subscribers to this theory accepted the premise that an anonymous composer had created every substantial folklore item at one point in time, much as a novelist produces a novel.

Folklore and fertility ritual

Various other origin theories have gained attention. The school of Cambridge University anthropologists, theorizing on the great comparative study by Sir James George Frazer, *The Golden Bough* (1st ed., 1890), converged on a central idea of myth-ritual origins of all folklore. In their view, a mythic narrative accompanied and explained a sacrificial fertility ritual among the heathens. In the course of time the myth becomes separated from the rite and floats independently in oral tradition, to splinter in turn into magic tales, popular ballads, nursery rhymes, and other folklore genres. Psychological folklorists ascribe the source of many folk narratives to dreams; the Hungarian Geza Roheim considers that dreams, precipitated by full bladders, engendered the widespread flood myth. Ballad scholars, such as Francis Gummere in *Old English Ballads* (1894) and other works, attributed the composition of ballads not to any single bard but to the joint efforts of a singing, dancing throng.

Origins by social classes have also formed the basis for widely held theories. One thesis, particularly identified with the German Hans Naumann and his term *gesunkenes Kulturgut* (literally, "downsinking cultural value"), contends that folklore originates with the aristocratic upper class, whose court poetry, mimes, bardic recitals, and pageants filter down in debased form to the peasantry. Russian and east European folklorists have sharply challenged this idea since the 1930s and substituted their own concept that folklore arises from the people, the folk, in expression of protest and outrage against the exploiting nobles and landowners. Hence in Russia folk bards and storytellers are honoured along with novelists and poets. In "A Theory for American Folklore" (1959), Richard M. Dorson has argued for a distinction between Old World and New World folklore origins. The folk traditions of North and South America combine the imported lore of the conquerors, the aboriginal lore of Indian tribes, and the lore arising since the colonizing period as a result of New World history and geography.

THE FORMS OF FOLKLORE

The elusive materials of folklore are best defined through the formal genres into which they fall. Genre definition has its own pitfalls, since attempts at neat categories invariably slice off related forms; however, folklorists do agree on certain broad kinds of traditions. These may be divided into oral literature, custom and festival, and material culture.

Genres of oral literature

Oral literature. The genres of oral literature cover spoken and sung expression. They may be further divided into the two large groupings of folk narrative and folk song, and such other smaller genres as proverbs, riddles, and beliefs or superstitions. *Folk narrative* is an umbrella for a wide range of oral prose traditions. Among them can be mentioned the *myth*, a semisacred adventure of a god or demigod set in the remote past; the *Märchen* or *fairy tale* (also *wonder tale* or *magic tale* and sometimes just *folktale* or *tale*), a pan-European popular fiction with aristocratic characters, magical episodes, and a symmetrical structure; the *legend*, a believed report often told conversationally and allusively; the *saga*, a personal, family, or local chronicle of marvellous oral history; the *romance*, a lengthy, adventure-filled narration with realistic characters; the *noodle* or *numskull tale*, relating the comical stupidities of a foolish person or a village of fools; the *jest* or *joke*, a short humorous fiction, often obscene and usually climaxed with a punch line; the *anecdote*, a brief traditional incident concerning a laughable action or saying of a historical personality; the *animal tale*, characterized by talking animals with human traits; the *cantefable*, a story containing songs or rhymes; and still other forms. Folk song, too, embraces myriad species. An important aspect of all folk songs is their association of text with tune, requiring the folklorist to trace the melodic as well as the textual history. A major division separates the *ballad*, which embodies a narrative, and the *lyric*, which expresses emotion, although, like *Märchen* and legend, the two basic forms often coalesce. The ballad can be further subdivided into the *Child ballad*, named for Francis James Child, who assembled 305 basic types of the English and Scottish traditional ballad (1882–98); the *broadside ballad*, a later development of the 16th and 17th centuries when balladmongers wrote up sensational events on broadsheets or broadsides and hawked them in the streets; and *European* and *American ballads*, which have evolved from local events. Other folk song genres range from the Russian *bylina*, or epic songs of a medieval hero, to the *lullaby*, used to sing a child to sleep.

Still another category of oral literature comprises *folk speech*, as distinct from formal or standard speech, and various traditional kinds of expressive utterances. Prominent among them are the *proverb* or folk saying, embodying wisdom in pithy phrases; the *riddle*, an enigmatic question paired with a deceptive answer; the *tongue twister*, a nonsense sentence difficult to pronounce because of its string of assonances; the *toast*, a convivial expression voiced as a drinking salutation; along with other forms involving a special use of language. *Beliefs* or *superstitions* are sometimes expressed as wise sayings, although they may also appear in tales and in customs.

Material culture. At the opposite pole from oral literature in the spectrum of folklore lies material culture or folk life, terms used to denote the physical objects produced in traditional ways. Material culture thus embraces folk architecture, folk arts, and folk crafts. Under these headings can be placed the construction of houses, the design and decoration of buildings and utensils, and the performance of home industries, according to traditional styles and methods. The shape of fences, the making of sorghum molasses, and the sewing of quilts all fall under material culture.

Custom and festival. Between oral and physical folklore there is a large middle ground filled by custom, ritual, festival, children's games (believed to be versions of adult rituals), folk drama, play parties, rites of passage, folk dances, and equivalent genres involving action, performance, and paraphernalia. To the verbal and tangible elements are added group behavioral traits.

The functions of these three genres vary for individuals and societies, but generally it may be said that material culture fills economic and aesthetic functions, that oral literature fills didactic, recreational, and educational functions, and that custom and festival function to provide psychic reassurance against external dangers. Rites placate gods and demons, tales and songs entertain and instruct, and home-baked bread on a hand-carved table fills the stomach while pleasing the eye.

THE GENERAL FUNCTIONS OF FOLKLORE

The escapist function

A behavioral classification of functions. Oral literature, like written literature but even more pronouncedly, satisfies the desires of mortals to transcend their mundane world. In the myths the heroes visit the otherworld; in the legends ordinary folk are taken to fairyland; and in the *Märchen* the youngest son or daughter of a peasant family makes a royal marriage. Gold, treasure, and wealth are prominent in narrative folklore. Jason and the Argonauts in Greek myth set out for the Golden Fleece; in fairy tales a magic goose lays golden eggs; and in real life, treasure seekers dig for the buried booty of pirates and outlaws. The Irish tradition of "Seán Palmer's Voyage to America with the Fairies" (in *Folktales of Ireland* by Sean O'Sullivan), told as an actual experience befalling Seán Palmer of County Kerry, combines the traditional theme of fairies transporting mortals to distant cities with a wish fulfillment fantasy. Seán believes the fairies take him by boat overnight to the United States, where he visits relatives in New York and Boston and is given money, new clothes, and tobacco. Many Irishmen have relatives in the U.S. whom they consider rich and long to visit.

Another way in which folklore lifts man above his narrow confines is in the breaking of powerful social taboos. What man can never do in actuality without incurring severe punishment the trickster in black American narratives, for example, does with impunity, such as John the slave slapping the face of his white mistress. Modern jokes with their heavy emphasis on genital-anal humour similarly arouse merriment by flouting taboos.

The etiological function

Mythic narratives and explanatory episodes in other kinds of folktales answer man's perennial questions: How did the world begin? Who created man, the animals, and the plants? The book of Genesis in the Bible gives two versions of a creation myth with worldwide analogies. Myths of origin are always ethnocentric; they explain the creation, by supernatural powers, of the people possessing the myth, a chosen people. The hostile, formidable, mysterious universe assumes a familiar outline through the personification of a creator and his adversary and of associated deities and demons with their own spheres of influence. Tribal peoples need their myths to safeguard their identity. An Ojibwa Indian in northern Michigan asked a researcher on their first meeting, "Do they have thunderstorms in South America?" On receiving an affirmative answer the Ojibwa turned morose and taciturn and refused to speak further. He had wished reassurance that thunderstorms existed only in the neighbourhood of Lake Michigan, since the Ojibwa protective deities were thunder spirits. Later, when mollified, he related a version of the creation myth, synthesized with the flood myth. The culture hero-trickster figure, Winabijou, flees from the underneath serpents, who send flood waters after him. He takes refuge atop a pine tree and barely keeps his head above the water that covers the earth. Water animals swim to him, and he sends them down to the bottom; the muskrat comes up with mud in its paws, with which Winabijou makes a little island. Each day the island gets bigger, and Winabijou sends the fox along the shore to circle it and report back. Eventually the island grows to the size of the present earth. "And you can see a fox trotting along the shore today," the narrator concluded. "The shore gets bigger too. Sand multiplies. When I first came here the beach on the cove was only half as big as it is now." The myth satisfactorily accounts for his universe.

The justifying function

Just as myths are generated to explain the physical world and its inhabitants, so does folklore at large provide support for the institutions and behaviour patterns of a culture. In nonliterate societies the central events of human life—birth, attainment of manhood, marriage, and death—are invested with elaborate ceremonials by which the social organism marks the mortal's progress and final exit to the afterlife. In modern society the quasi-religious character of these passage rites, as Arnold van Gennep has termed them, can be observed in the rituals of baptism, confirmation or graduation, weddings, and funerals, which are sometimes secular, sometimes ecclesiastical, and sometimes a mixture of both. In contrast to supernatural beings, who acquire anthropomorphic personalities and walk the earth (like Winabijou, or Christ and St. Peter), historical heroes acquire supernatural attributes and move heavenward. The illustrious secular heroes and heroines—like Joan of Arc, Confucius, Peter the Great, Julius Caesar, King Alfred, and George Washington—become national symbols, consecrated with monuments, holidays, hagiographic literature, and folk legends. They function in the national culture as standard-bearers of the values and goals of the nation. The legend of young George Washington saying to his father, "I cannot tell a lie, it was I who chopped down the cherry tree," drives home the moral of the proverb ascribed to Benjamin Franklin (but actually common in Europe in the 17th century and earlier), another culture hero: "Honesty is the best policy."

The pedagogic function

A good deal of folklore, especially the proverbs, fables, cautionary tales, and confessional ballads, serves to instruct and remind members of society of wise codes of conduct. The proverb and the fable may teach the same lesson; "Pride goes before a fall" is illustrated in Aesop's fable of the cat who praised the crow's singing voice, so that the crow up in the tree started to sing and dropped the cheese from her mouth to the ground. Proverbs may sometimes appear contradictory. "Look before you leap" is challenged by "He who hesitates is lost," but each saying contains its truth to be applied to a given situation. A whole ethic may be summarized in a proverb, as is the Protestant ethic of hard work in the saying, "Early to bed, early to rise, makes a man healthy, wealthy, and wise."

Although proverbial expressions are still frequently uttered in industrialized societies, they are used as mild commentary and poetic cliché rather than, as in African societies, for firm codes with the force of judicial precedent. The African novelist Chinua Achebe writes in *Things Fall Apart,* "Among the Ibo the art of conversation is regarded very highly, and proverbs are the palm-oil with which words are eaten." And his novels are sprinkled with proverbs. Actual court cases among the Anang Ibibio of Nigeria in which judges were swayed by proverbs have been recorded by the anthropologist John Messenger. On one occasion the chief judge advised the plaintiff and his witnesses, "If you visit the home of the toads, stoop," when they refused to make their statements under oath and thereby forfeited their case. The equivalent precept, still forceful in modern life, is "When in Rome do as the Romans do." Among the Bambara-speaking people of Mali in Africa, the *griots,* the celebrated epic bards, customarily "warm up" before their recitations and during breaks in their lengthy performances by singing proverbs in rapid sequence. This practice serves the function of gaining the attention and respect of the audience, who think of proverb sayers as wise men knowing how the society works; hence the listeners will be ready to credit the historical tradition that follows.

The controlling function

Besides suggesting rules for conduct, folklore also drives home the need for proper social behaviour by holding up to scorn those who depart from socially accepted norms and by eulogizing those who follow them. Jokes, the most popular form of oral literature in modern society, ridicule stereotyped characters who display traits disparaged by the Establishment. The stingy Scotsman, the miserly Jew, the ignorant Irishman, and the stupid Polack are all caricatured in joke cycles whose implicit values laud the qualities of liberality, generosity, intelligence, and cleanliness. Anecdotes of local characters similarly excoriate the slovenly, shiftless, degenerate, gullible, and naïve. One favourite theme of anecdotes is the lazy man, often identified with some local ne'er-do-well. Starving to death, the lazy man is offered popcorn: "Is it shelled?" he asks, and prefers to starve rather than do the labour of the shelling. Just as proverbs teach the gospel of work, so do such anecdotal legends mock the idle and improvident. Conversely, folklore also exalts individuals who exemplify the virtues considered admirable in the culture. Saints' legends recount the miraculous cures and rescues that the saint revered by the folk (who is not necessarily canonized by the church) has effected for the faithful. Devout members of the Mormon Church continue to relate experiences of succour given them in distress by one of the three Nephites, who appears as a stranger in white garments to render assistance and vanishes as mysteriously as he materializes. Heroes in ballad, legend, and *Märchen* reflect the dominant values of their societies and are rewarded by success. The champion of heroic epics is an invincible warrior. The hero of Negro ballads is a big-talking badman. The modern hero in the anecdotal legends of youthful dissenters is the antihero who sells LSD pills or marijuana and outwits narcotic agents, the police, and draft boards. All these heroes symbolize ideal types for their social groups. Deviation from the ideal brands the nonconformist a coward, an Uncle Tom, or an Establishment toady. In the form of folklore known to 17th-century American Puritans as the "remarkable providence," God punished sinners, blasphemers, heretics, witches, marauding Indians, and critics of the state with acts of supernatural vengeance, while rewarding his elect with supernatural assistance during their wilderness trials. Puritan clergy collected and published these providences as lessons and guideposts for their people. In Socialist countries of the 20th century, the governments have taken a hand in the process of regulating conformity by awarding prizes to folk poets and folk narrators who attack bourgeois and capitalist villains and extol peasant and Socialist heroes.

A cultural-geographical classification of functions. Generalizations about the functions of folklore need to be qualified by considerations of culture area and social milieu. One kind of distinction should be made between Old World and New World folk traditions, since North and South America were colonized by Europeans, who then transported African slaves to their overseas empires. Consequently, a given country of the New World possesses several coexisting and interacting traditions; the indigenous Indian; the colonizing Spanish, Portuguese, French, or English; the African Negro; the 19th- and 20th-century European and Asian immigrant; and the national and regional, shaped by new historical and environmental factors. The continent of Australia also belongs with the New World in this regard, its folklore being divided between the Aborigines and the English settlers. In the colonized nations, each folk tradition serves to reinforce the group identity of its members in a pluralistic culture. Syncretism, or the fusion of different traditions, is a process especially characteristic of New World folklore. The miraculous appearance of the Virgin of Guadalupe to the Aztec Indian Juan Diego in 1531, who received from her a painting of herself on his cape, brought together Spanish-Catholic and pre-Conquest Indian strains. Mexican Indians and mestizos identified the Virgin with an Aztec goddess Tonantzin, pictured her as dark in hue, and celebrated her with costumed dances of pagan origin but within the framework of Roman Catholic worship and beatification. The ballad of "John Henry," relating the contest between a brawny Negro labourer and a new steam drill on a railroad tunnel in West Virginia in the 1860s, is widely sung by both American blacks and whites. John Henry dies "with his hammer in his hand," and interpreters see him as a symbol of the unequal struggle of the black man against the white, and of man against the machine. Immigration has brought large numbers of Slavs to western Canada, Japanese to Brazil, Italians to Argentina, Germans to Chile, and many peoples to the United States, whose folkways partially survive the transoceanic crossing. The traditional lore physically connected with the original homeland seems to vanish the most quickly; thus the Irish belief in fairies does not survive, since the grassy circular "fairy rings" inhabited by the fairies of the Old Country are not found in America.

Literate and nonliterate functions

Another large distinction in the functioning of folklore separates the so-called nonliterate societies lacking written languages from the advanced civilizations engulfed in print. Reliance on the oral tradition is obviously far greater in the nonliterate cultures. The matter of recording history provides one example: instead of referring to printed books containing a great many facts that no one person could or would wish to keep in his head, the tribal groups depend upon reciters of historical traditions, some of whom, like the *whare wananga* of the New Zealand Maori, the sagamen of 11th-century Iceland, and the *griots* of Mali, have a professional status. These annals and genealogies bear little resemblance to documented written histories and belong properly to the products of oral literature, since they include magical and marvellous episodes of folktales and legends. They do correspond in function to national histories in celebrating the achievements and culture heroes of a chosen people. An ingenious study by two teams of American anthropologists who collected traditions from two rival Indian tribes in northern California revealed a similar patterning but contrasting roles for the same characters and events. The enemy tribe was always guilty of the provocative incident, fought dishonourably, and suffered defeat.

Even in literate civilizations large enclaves of largely oral, tradition-directed cultures persist. Narrators with enormous repertoires have been located in the 20th century in the Gaelic-speaking highlands and islands of western Scotland and western Ireland. Peig Sayers of the Great Blasket Island related 375 tales to the collectors of the Irish Folklore Commission, many of them long wonder tales taking all evening or several evenings to complete. In modern urban, technological society, such reciters find no audiences. Life has too many distractions and diversions and sources of information. Nevertheless, story-telling continues among the highly literate, in the form of joke-telling sessions, which may last an hour or two; but the jocular narratives are relatively brief, snappy, and quickly climactic. They serve as icebreakers to establish camaraderie in a group, although they may also give offense to some members of a social gathering with differing political or moral views. Ballad making originally made possible the dissemination of sensational news in attractive form and receded with the advent of newspapers, which emulated many of the same techniques of ballad makers in their handling of lurid news items.

Correspondence between oral and literary cultures

Oral and literary cultures are no longer sharply divided. On the lower levels of society cheap printed sources—such as chapbooks, broadsides, jestbooks, mass magazines, almanacs, and newspapers—feed materials into and draw from oral streams of lore. In the electronic age, according to Marshall McLuhan, the literate world is reverting to an oral-aural community.

Folklorists at first identified their subject with the rural peasantry. Seasonal festivals, old wives' remedies, supernatural beliefs in demonic figures, and magic makers flourished in the countryside. Cities were the centres of learning, industry, and wealth, and the spoilers of tradition and folklife. This easy contrast is being revised as folklore scholars turn their attention to the cities where ethnic neighbourhoods and occupational workers form closely knit, tradition-bound societies. A team project of Hungarian folklorists investigating Budapest reported many vigorous manifestations of traditional behaviour, such as May Day celebrations, dancing parties held by tradesmen, housewarming ceremonies, and occupational jokes. In a detailed study of a Tokyo district, R.P. Dore reported on households that maintain traditional festival practices and the ancient Shintō beliefs in *kami* (deities) and *fuda* (shrine amulets). But city life does of course affect traditional ways brought in from the country. The ubiquitous Japanese-peasant conception of the *Kappa*, a dangerous boy-goblin with an inverted saucer on his head containing magical water, is found in Tokyo in newspaper comic strips, on decorative designs, or as a character in a novel by Akutagawa Ryūnosuke. In a series of ethnographic studies in Yucatán, the American anthropologist Robert Redfield compared four communities, ranging from the simple village to the complex town, and found that living rituals in the village become desiccated superstitions in the town.

Field research in the steel-manufacturing cities of Gary and East Chicago, Indiana, by Indiana University folklorists has shown not only a decrease in the folklore genres of the many ethnic groups living there but also an emergence of a new city lore. The chief elements of this lore derive from the human side of the all-encompassing steel industry, from fears and rumours about crime, and from anecdotes and stereotypes about other racial and ethnic groups. This emerging lore provides a sense of solidarity and rootedness in the impersonal, unlovely city.

THE DIFFUSION AND ACCULTURATION OF FOLKLORE

Folklore is at once remarkably stable and remarkably shifting. These phenomena have intrigued folklorists, who explain the apparent paradox through the mechanism of diffusion. A complex item of folk tradition, whether an epic song, a house style, a folk drama, or a wonder tale, once it has attained coherent form, may travel across continents and oceans and endure through the centuries. The basic type retains its outline but external features are continuously modified and adapted to new surroundings. Barriers of language, religion, and culture offer no obstacles to the movements of folk products. The most bitter of enemies share the same traditions. Supporters of diffusion challenged the upholders of independent invention in the latter decades of the 19th century and eventually won the day as depth field studies proved clearly the wanderings and migrations of individual texts. The first comparative study of a folktale, Marian Cox's *Cinderella* (1893), brought together 345 variants of the same recognizable story plot. Even legends, seemingly so anchored in time and place to specific historical events and personalities, were shown to migrate and become attached to different localities and heroes. The legend of the Returning Hero, sleeping in the fastnesses of the mountain with his warriors until his people would summon him to their rescue, fastened on to Frederick Barbarossa, King Arthur, and Thomas the Rhymer and is still told of modern figures. How did folklore diffuse? Proponents of diffusion offered various explanations: sailors, merchants, and travellers transported folk matter along the great historic trade routes; Gypsies acted as peddlers of folklore; borders between peoples served as bilingual zones through which songs, tales, and styles easily moved; imperial powers planted the traditions of the mother country in distant places. All these hypotheses have some merit; European tales found among North American Indians, for instance, undoubtedly date back to the period of their first culture contact with French and Spanish missionaries, soldiers, and traders. Yet Indian tales did not lodge in the narrative repertoires of the European settlers. The flow appears to go from literate to nonliterate cultures.

Acculturation represents a somewhat different phenomenon of folklore movement. Here not the lore but the folk move, bearing with them their ancestral heritage. What happens to this heritage when peasants of India resettle in Africa, when Chinese move to Malaysia, when European communities pour into the New World? The resulting process of adjustment, adaptation, compromise, and assimilation is called acculturation. But what takes place in this process is not so easily understood. Parts of the Old Country tradition recede into "memory culture," to be recalled on direct stimuli but otherwise lying dormant in the recollections of the first generation. Other parts remain in vigorous use. Still other parts take on New World coloration. Among Greek immigrants in the United States, for instance, the telling of paramythia (or *Märchen*) and the sighting of creatures like the *vrykólakas*, a vampire demon, have virtually ended, but the belief in the power of the evil eye and of the protective power of saints remains as strong as ever. As a general principle, the beings of lower mythology do not transplant into an alien environment, whereas the daily and seasonal rituals and associated beliefs can be maintained within compactly settled groups. The rate of acculturation will vary greatly between immigrant societies, depending upon what conservative forces are at work. Such groups as the Pennsylvania German Amish and the Hasidim in Brooklyn, N.Y., retain a high degree of their magical belief system, folk arts, and ritual observances, because they have made strenuous efforts to wall themselves in from the mainstream of American culture, an effort strengthened by their independent churches. By contrast, Irish-Americans after a century in their new homeland have lost most of their distinctive folklore and blended in with the rest of the general Catholic and middle class population.

The mobility and technology of modern man have noticeably affected the diffusion and acculturation of folklore. People emigrate by airplane, communicate by telephone, and soak information from the mass media. Yet old ways of thinking and believing show remarkable tenacity. Many folk today discredit the landing of men on the moon as a publicity hoax contrived by the United States government and the television networks. These are the people who have seen flying saucers, been cured by faith healers, and conversed with revenants. Folklore is as old, and as young, as man. (R.M.D./Ed.)

Folk art: general considerations

19th-century notions of folk art

Under the influences of 19th-century sentiments, the concept of "folk" was affected by a fondness for the picturesque, by generous impulses toward aesthetic democracy, by the so-called arts and crafts movement, by historicism and philosophical Idealism, by nationalism and regionalism, and by an occasionally aggressive, wildly mystical sort of ethnocentrism. Thus, folk art was defined, if only by the implications of usage, sometimes as almost anything that could be considered quaint; sometimes as all non-elite art, primitive and popular included; sometimes as whatever seemed vaguely homemade; sometimes as the art of the common people, but with the latter regarded as a ghostly entity existing outside of real class structures; and sometimes as a traditional, characteristic art that preserved a cultural heritage and somehow represented the collective soul of a nation or a province. There are some more or less acceptable elements in such definitions. What is missing is a flat enough acknowledgment that folk art is the recognizable product of a nonruling, relatively unaffluent social class and that references to it except in relation to the elite art of a ruling class are meaningless.

A core description of folk art

Putting all these considerations together, one arrives at what can serve as a core description. Unmistakable folk art is the art of a class of peasants, herdsmen, seamen, artisans, and small tradesmen living as a rule away from urban centres in societies that, while literate and highly civilized, are not highly industrialized. Such societies were to be found in Europe, particularly eastern Europe, until well into the 20th century, and they will still exist, of course, on other continents. Where they no longer exist, the term folk may still be appropriate for the art of social or ethnic minorities that have preserved earlier traits by living apart, culturally and often physically. More doubtfully, the term may be extended to certain traditional artistic productions—especially those associated with carnivals, national holidays, religious processions, and the like—that are the work of thoroughly unfolkish city people.

FUNDAMENTAL CHARACTERISTICS OF FOLK ART AND THE FOLK ARTIST

Although the folk kind of art is invariably produced, according to the only definition that makes serviceable sense, in a culture that also produces the elite kind for a ruling class, the usual elite classification of the arts into the fine, or independent, and the useful, or decorative or dependent, is pretty much ignored by folk visual artists, who are typically willing to devote their creative imaginations to the design or decoration of tools, toys, furniture, cottages, clothes, arms, banners, musical instruments, and so on. The portrait of a Spanish patron saint may appear in embroidery; narrative pictures belonging, according to the European academic classification, to the elevated "history" category are painted on Sicilian carts. Something of the same mixture of the so-called high and low is typical of folk literature, music, and dancing: tragedy takes ballad form and a noble war dance the form of a Highland fling.

Artists and patrons. In some ways the usual folk artist can be called an amateur. He lacks professional training, he normally has a second, more remunerative occupation, and he creates primarily for his own pleasure and needs or at most for those of acquaintances in his own small locality. But in other ways he is quite unlike an amateur artist in a nonfolkish social context. He is apt to work strictly within the limits of a traditional formula, he often remains anonymous, and often he is an artisan who carries over into his works of art his everyday skill. Amateur is not quite the word for a blacksmith who designs a fantastic shop sign in wrought iron nor for a baker who turns out figurines.

Materials and techniques. Works of visual folk art are rarely executed in anything except readily available, fairly inexpensive materials. A sculptor is likely to work not in marble or bronze but, rather, in wood, iron, clay, straw, ice, sugar, or whatever else is at hand; a painter may substitute an old board for the elite artist's canvas or special paper. Perhaps because of this situation and perhaps because of the already mentioned tendency to ignore the separateness of fine art, the folkish creator displays little of the elite (or, for different reasons, the primitive) concern to employ a specific technique for a specific material—to respect what a French art historian has called the vocation of each sort of artistic material. Folk wooden sculpture may be carved in a manner suitable for granite; painting may be done in an approximation of embroidery technique; pottery may bear incised patterns borrowed from metalwork or basketry; and almost any material is apt to be disguised under bright colours.

Styles. Encouraged by local patriotism, scholars have divided folk art into mostly national and regional styles: English, French, German, or Spanish folk music, Breton, Provençal, or Balkan costumes, Pennsylvania Dutch decorations, Mexican processional figures, and so on. This sort of classification is justified by the remarkable traditional and settled nature of much of folk art. But it should not be allowed to obscure, as it frequently does, the existence of other sorts of style, nor should it suggest that these national and regional manners are without histories. There are Jewish, Christian, Islāmic, and other religious folk styles; mountain, valley, and seaboard styles; classical, romantic, realistic, fantastic, and expressionistic styles; and naturalistic and abstract styles. There are period styles; in fact, a sizable number of vaguely ahistorical-looking provincial folk styles are arrested styles of the 18th and early 19th centuries.

MEANINGS AND FUNCTIONS OF THE FOLK ARTS

Conservative and preservative qualities in folk art

Although the folk, defined as a class, live in nonprimitive societies, they are more or less isolated from the urban centres in which cultural change may be relatively rapid; and in their predominantly rural environments they are close to nature and to the basic birth-to-death cycle of human existence. Hence folk art is strongly conservative and preservative; in fact, it is often referred to as simply traditional art (with the unfortunate implication that other kinds of art do not possess traditions). It acts as a repository for proverbial wisdom, ancient superstitions, sentimental themes, and religious beliefs that may have long since ceased to be orthodox; it accompanies and celebrates what is repetitive in the lives of individuals and of communities—baptisms, marriages, funerals, anniversaries, sowings, reapings, and the daily routine of work. Often it preserves customs the meanings of which have been largely forgotten; the colouring of Easter eggs and the decorating of Christmas trees are familiar examples. Often, too, the old meanings are given an unexpected twist by a naïve or eccentric artist. But, on the whole, folk art functions within its sphere, as does primitive art in an entire society, on the side of continuity and social stability; and there is some justification for 19th-century talk about the expression of a collective "soul."

The "soul," however, is not so much that of a nation, a race, or a province as that of a social and economic class. The best folk art reflects the common sense, unpretentiousness, humour, and peculiar slyness of humble people in contact with small, hard realities; and not infrequently it reflects hostility toward the rich and powerful. In folk depictions of the Nativity, a gift bearer may bring not frankincense and myrrh but, instead, a jacket for the baby Jesus; in a harvest song, one may suddenly be informed that "the gentlemen," not the corn, are about to be mowed.

THEORETICAL ISSUES

Diffusionism, environmentalism, and related theories have aroused polemicists in the field of folk art as well as those in primitive art. But perhaps stronger feelings can be, or have been, stirred among folk-art specialists by theories that attempt to answer the two rather loaded questions of how folk art is created and, second, what the difference is between a folk art and a folkcraft.

Theories about the creation of folk art

The first question has some of its roots in the facts that the adjective folk may connote "communal" and that folk art is often the work of unknown creators. Other, probably more important, roots lie in the fact that, by the beginning of the 19th century, German Romantic thinkers were committed to myths (at that time fairly harmless) about the collective spirit, or virtue, of a nation or a race. The result was the idea, set forth with varying degrees of explicitness, that folk art—in particular, poems, melodies, and dances—was a communal creation. This idea was eventually replaced by a theory of communal re-creation or communal growth in situations in which the work of an individual poet, singer, or dancer is preserved only in memories and is modified as it passes from locality to locality and from generation to generation. In the meantime another largely German idea, that of cultural sinking, was provoking argument. According to this idea, folk art—visual art in particular—is derived, after a lapse of several decades or even several centuries, from the elite art of a ruling class; thus a Crucifixion carved by a 19th-century Breton peasant is explained as a late-provincial imitation of the 17th-century Baroque of Paris and Versailles. On all these matters the opinion of a majority of modern scholars is apt to be conciliatory. Although collective improvisation is not impossible, the original creator of a work of folk art is practically always a single artist. But it is possible to grant that a good deal of communal re-creation occurs in literature, music, and the dance and that in the folk visual arts the role of the individual creator, subject as he is to the demands of tradition and function, is not as decisive as it is in elite art. Again, it can be granted that folk art is often a belated borrowing from elite. But this need not prevent recognition of the fact that the folk as a class have their own traditional themes and forms.

Difference between a folk art and a folkcraft

The question about the difference between a folk art and a folkcraft is partly philosophical (which is not a reason for neglecting it) and partly quite practical, at least for museum curators who must classify exhibits, for specialists interested in a proper division of labour, and for modern educators or animators who want to continue the 19th-century arts and crafts movement. It has been argued that, where an artist creates, a craftsman merely fabricates; that, whereas the first is a nonmaterialist who makes something for the sake of expression or of disinterested contemplation, the second is a materialist who makes something for a definite function; and that, whereas the first cannot know exactly what he is going to produce, the second has an extremely precise foreknowledge accompanied by a plan and a method of execution. On this argument many alleged works of folk art can be dismissed as works of mere craft. But the argument is challengeable, partly on the familiar ground that art is in the eye of the beholder.

Although folk art, as was pointed out above, may borrow from elite and although it may also lend motifs to everybody from classical musicians to modern dress designers, it has proved incapable of surviving in the world of 20th-century industry and communications. Like primitive art, it can live only in isolation. Unlike primitive, however, it has demonstrated, particularly in the performing arts, an extraordinary ability to transform itself into a special branch of popular art (see POPULAR ARTS).

(R.McMu./Ed.)

Folk literature

Folk literature, as mentioned in the introduction, may also be called folklore. It is, in fact, the lore (traditional knowledge and beliefs) of cultures having no written language, and is, by definition, transmitted by word of mouth. It consists, as does written literature, of both prose and verse narratives, poems and songs, myths, dramas, rituals, proverbs, riddles, and the like. Nearly all known peoples, now or in the past, have produced it.

Until about 4000 BC all literature was oral; but beginning in the years between 4000 and 3000 BC writing developed both in Egypt and in the Mesopotamian civilization at Sumer. From that time on there are records not only of practical matters such as law and business but increasingly of written literature. As the area in which the habitual use of writing extended over Asia, North Africa, and the Mediterranean lands and eventually over much of the whole world, a rapid growth in the composition of written literature occurred, so that in certain parts of the world, literature in writing has to a large extent become the normal form of expression for storytellers and poets.

Nevertheless, during all the centuries in which the world has learned to use writing, there has existed, side by side with the growing written record, a large and important activity carried on by those actually unlettered, and those not much accustomed to reading and writing.

ORIGINS AND DEVELOPMENT

Of the origins of folk literature, as of the origins of human language, there is no way of knowing. None of the literature available today is primitive in any sense, and only the present-day results can be observed of practices extending over many thousands of years. Speculations therefore can only concern such human needs as may give rise to oral literature, not to its ultimate origin.

The nature of oral traditions. Nor can any evolution in folk literature or any overall developments be spoken of explicitly. Each group of people, no matter how small or large, has handled its folk literature in its own way. Depending as it does upon the transmission from person to person and being subject to the skill or the lack of skill of those who pass it on and to the many influences, physical or social, that consciously or unconsciously affect a tradition, what may be observed is a history of continual change. An item of folk literature sometimes shows relative stability and sometimes undergoes drastic transformations. If these changes are looked at from a modern Western point of view, ethnocentric judgments can be made as to whether they are on the whole favourable or unfavourable. But it must be remembered that the folk listening to or participating in its oral literature have completely different standards from those of their interpreters.

Nevertheless, two directions in this continually changing human movement may be observed. Occasionally a talented singer or taleteller, or perhaps a group of them, may develop techniques that result in an improvement over the course of time from any point of view and in the actual development of a new literary form. On the other hand, many items of folk literature, because of historic movements or overwhelming foreign influences or the mere lack of skillful practitioners of the tradition, become less and less important, and occasionally die out from the oral repertory. The details of such changes have been of great interest to all students of folk literature.

Early development

The beginnings of written literature in Sumer and Egypt 5,000 or 6,000 years ago took place in a world that knew only folk literature. During the millennia since then written literature has been surrounded and sometimes all but overwhelmed by the humbler activity of the unlettered. The emergence of the author and his carefully preserved manuscript came about slowly and uncertainly, and only in a few places initially—the literary authorship that flourished in the Athens of Pericles or the Jerusalem of the Old Testament represented only a very small part of the world of their time. Nearly everywhere else the oral storyteller or epic singer was dominant, and all of what is called literary expression was carried in the memory of the folk, and especially of gifted narrators.

All societies have produced some men and women of great natural endowments—shamans, priests, rulers, and warriors—and from these has come the greatest stimulus everywhere toward producing and listening to myths, tales, and songs. To these the common man has listened to such effect that sometimes he himself has become a bard. And kings and councillors, still without benefit of writing, have sat enthralled as he entertained them at their banquets.

Cultural exchange in written and oral traditions. This folk literature has affected the later written word profoundly. The Homeric poems, undoubtedly oral in origin and retaining many of the usual characteristics of folk literature, such as long repetitions and formulistic expressions, have come so far in their development that they move with ease within a uniform and difficult poetic form, have constructed elaborate and fairly consistent plots and successfully carried them through, and have preserved in definitive form a conception of the Olympic pantheon with its gods and heroes, which became a part of ancient Greek thinking.

Not everywhere has the oral literature impinged so directly on the written as in the works of Homer, which almost presents a transition from the preliterate to the literate world. But many folktales have found their place in literature. The medieval romances, especially the Breton lays, drew freely on these folk sources, sometimes directly. It is often hard to decide whether a tale has been learned from folk sources or whether a literary story has gone the other way and, having been heard from priest or teacher or doctor, has entered oral tradition and has been treated like any other folktale or folk song. The unlettered make no distinctions as to origins.

Influence on modern literature

As the European Middle Ages lead into the Renaissance the influence of folk literature on the work of writers increases in importance, so that it is sometimes difficult to draw a sharp line of distinction between them. In literary forms such as the *fabliaux* there are many anecdotes that may have come ultimately from tales current among unlettered storytellers, but these have usually been reworked by writers, some of them belonging in the main stream of literature, like Boccaccio or Chaucer. Only later, in the 16th and 17th centuries, in such works as those of Straparola and Giambattista Basile, did writers go directly to folk literature itself for much of their material.

Since classical times composers of written literature have borrowed tales and motifs from oral narratives, and their folk origin has been forgotten. Examples abound in Homer and *Beowulf.* In their literary form these stories have often lived on side by side with tellings and retellings by oral storytellers. Modern examples of traditions so used are found in Ibsen's *Peer Gynt* and Gerhart Hauptmann's *The Sunken Bell.* Particularly frequent in all literature are proverbs, many of them certainly of folk origin.

In Finland a good example of the direct use of folk literature in the construction of a literary epic is seen in the *Kalevala,* composed by Elias Lönnrot in the 1830s, primarily by fusing epic songs that he had recorded from Finnish singers. The *Kalevala* itself is a national literary monument, but the songs Lönnrot heard are a part of folk literature.

Writers and songmakers have always used themes taken from oral legends and folk songs and in their turn have affected the traditions themselves. In recent years the cinema has presented old folktales to an appreciative public, and interest in folk songs especially has been stimulated by the radio and television. Inevitably this oral literature has become less truly oral, and much pseudofolk literature has been presented to the public, habituated as it is to the usual literary conventions.

Within urbanized Western culture it is clear that folk literature has been gradually displaced by books and newspapers, radio, and television. Persons interested in hearing authentic oral tales, traditions, or songs must make special efforts to discover them. There still exist isolated groups that carry on such traditions—old people, recent immigrant enclaves in cities, and other minority populations, rural or urban. Children are also important for the carrying on of certain kinds of oral traditions such as singing games, riddles, and dance songs. These go on from generation to generation and are added to continually, always within an oral tradition.

During the past few generations folk festivals have flourished. These have become almost worldwide and of the greatest variety. They are likely to revive older dances or bring in new ones from other countries, but they also have some singing and occasionally taletelling. Usually a genuine attempt is made to keep them within the authentic local tradition, and they have been a stimulus to the preservation of a disappearing phase of modern life.

If folk literature is actually dying out, the process is very slow. It is now, as it has always been, the normal literary expression for the unlettered of all continents.

CHARACTERISTICS OF FOLK LITERATURE

The most obvious characteristic of folk literature is the fact that it is oral. In spite of certain borderline cases it normally stands in direct contrast with written literature. The latter exists in manuscripts and books and may be preserved exactly as the author or authors left it, even though this may have happened centuries or even millennia ago. Through these manuscripts and books the thoughts and emotions and observations and even the fine nuances of style can be experienced without regard to time or distance. With oral literature this is not possible. It is concerned only with speaking and singing and with listening, thus depending upon the existence of a living culture to carry on a tradition. If any item of folk literature ceases to exist within the memory of man it is completely lost.

Effects of oral transmission

The speaker or singer is carrying on a tradition that he has learned from other speakers and he delivers it to a living audience. It may well be that his listeners have heard this material many times before and that it has a vigorous life in the community, and they will see to it that he does not depart too far from the tradition as they know it. If acceptable to the listeners, the story or song or proverb or riddle will be repeated over and over again as long as it appeals to men and women, even through the ages and over long geographical distances.

In some cultures nearly everyone can carry on these traditions, but some men and women are much more skillful than others and are listened to with greater pleasure. Whatever the nature of these tradition bearers, the continued existence of an item of oral literature depends upon memory. As it is passed on from one person to another it suffers changes from forgetting or from conscious additions or substitutions; in any case, the item changes continually.

The more skillful tradition bearers take pride in the exactness with which they transmit a tale or song just as they have heard it many years before, but they only deceive themselves, for every performance differs from every other one. The whole material is fluid and refuses to be stabilized in a definite form. The teller is likely to see places where he can make great improvements, and he may well begin a new tradition that will live as long as it appeals to other tellers. It thus happens that in nearly all cultures certain people specialize in remembering and repeating what they have heard. There are semiprofessional storytellers around whom large groups of people assemble in bazaars or before cottage fires or in leisure hours after labour. Some of these storytellers have prodigious memories and may with only slight variations carry on to a new generation hundreds of tales and traditions heard long ago.

Forms and functions

Certain bards and minstrels and songmakers develop special techniques of singing or of telling epic or heroic tales to the accompaniment of a harp or other musical instrument. In the course of time in various places special poetic forms have been perfected and passed on from bard to bard. Such must have been the way in which the remarkably skillful heroic meters of the Greek epics were developed.

A different kind of oral tradition is preserved by the ritual specialists: priests, shamans, and others who perform religious ceremonies and healing rites. Frequently these rituals must be remembered word for word and are not believed to be effective unless they are correctly performed. The ideal of such priestly transmitters of oral tradition is complete faithfulness to that which has been passed down to them.

Not least important of the many reasons for the existence and perpetuation of folk literature is the need for release from the boredom that comes on long sea voyages or in army camps or everywhere on long winter evenings. Some folk literature is primarily didactic and tries to convey to simple men the information they need to carry on their lives properly. Among some peoples the relation of man and the higher powers is of especial concern and gives rise to myths that try to clarify this relationship. Cooperative labour or marching is helped by rhythmic songs, and many aspects of social life give rise to various kinds of dance.

A great many of the special forms of literature now in manuscripts and books are paralleled in traditional oral literature, where history, drama, law, sermons, and exhortations of all kinds are found, as well as analogues of novels, stories, and lyric poems.

Relation to folklore and mythology

Folk literature is but a part of what is generally known as folklore: customs and beliefs, ritualistic behaviour, dances, folk music, and other nonliterary manifestations. These are often considered a part of the larger study of ethnology, but they are also the business of the folklorist. (For further discussion of this point, see above *Folklore as an academic discipline.*)

Of special importance is the relation of all kinds of folk literature to mythology. The stories of Maui and his confreres in the Pacific and of gods and heroes of African or American Indian groups have behind them a long and perhaps complicated history. This is especially true of the highly developed mythologies of India, and the Greek, Irish, and Germanic pantheons. All are the results of an indefinitely long past, of growth and outside influences, of religious cults and practices, and of the glorification of heroes. But whatever the historical, psychological, or religious motivations, the mythologies are a part of folk literature and, though traditional, have been subject to continual changes at the hands of the taletellers, singers of stories, or priestly conductors of cults. Eventually singers or storytellers of philosophical tendencies have systematized their mythologies and have created with fine imagination the figures of Zeus and his Olympic family and his semidivine heroic descendants. Though the details of these changes are beyond the scope of this article, stories of the gods and heroes and of supernatural origins and changes on the earth have played an important role in all folk literature.

TECHNIQUES OF FOLK LITERATURE

Since the tales, legends, and epic and lyric songs discussed here are a part of the experience of a preliterate group or at least of the essentially unlettered, they differ in many ways from literary works addressed to a reading public. Long forgotten are the person or persons originally responsible for the tradition that has resulted in examples of folk literature. Only the tale or song remains to be repeated and often changed by subsequent storytellers, singers, or bards. In the course of its history it is listened to by generations of the unlettered, and its success and its very survival depend on how well it satisfies their emotional needs and intellectual interests.

Since in essence all folk literature is oral and subject to its survival in the minds of men, it is full of devices to aid

Technique of repetition in folk tales

memory. Perhaps most common of all is mere repetition. Especially in folktales and epics it is common to hear the same episode repeated with little or no verbal change. As the hero encounters his successive adversaries the description changes only enough to indicate the increasing terror of the enemy, always leading to a climax and usually to the hero's success. These long repeated passages often enable the teller of tales or the singer of an epic to extend his performance as much as he desires.

Aside from repetition of entire episodes, folk literature of all kinds is filled with formulistic expressions. It may be the beginning or the ending of a folktale—the "once upon a time" or the "married and lived happily ever after" or sometimes quite meaningless expressions—or standard epithets attached to certain persons or places. These formulas are so characteristic of oral literature that an abundance of such commonplaces seems to be a guarantee of authentic oral origins even of a great epic.

These formulas are matters not only of words but of structure. The storyteller or singer has at his disposal a large variety of conventional motifs and episodes and may use them freely. How appropriately they are made a part of his composition depends on his skill, but his listeners are not likely to be very critical so long as he keeps them interested. Indeed it is remarkable that in spite of this apparent freedom of improvisation so many rather well-articulated plots have lived for centuries retaining all their essential features. It is this combination of a basic narrative type with a freedom of treatment within traditional limits that makes it possible to identify hundreds of versions of the same tale or song as they appear over long stretches of time and space.

Though much of narrative folk literature is frankly fictional and filled with unrealistic events, the successful storyteller or epic singer gives his story credibility by the use of realistic details. Often these are merely homely touches linking the never-never land of the tale or song to everyday life or emotions. For the unlettered listeners such realistic details may allow a stretching of the imagination to embrace a larger world. Heaven or hell it may be or kingly palaces where the peasant hero rules with a splendour only known to those who have never seen a court. Often these details are given only to ensure that willing suspension of disbelief characteristic of all fiction, but sometimes a realistic touch, even in the midst of weak motivation and violence, may give nobility to a mediocre tale or song.

Repetition, formulas both in words and in structure, realism enough to support the marvelous in tale or song, violent actions and simple strong emotions—these qualities are generally found in all folk literature. The varying demands of the listeners are all-important influences. In some cultures this implies that actions should be well motivated so that listeners may identify themselves with certain characters. But in others, such as in many parts of India and in many preliterate cultures, motivation is often weak or entirely lacking.

For lyric songs, proverbs, riddles, and charms (and often legends), the relation of artist and audience is of little importance.

REGIONAL AND ETHNIC MANIFESTATIONS

Ethnic variations in relating folktales

In many particulars of form and substance there will be found great variations in the ways folk literature is manifested. The interests of people in one culture may differ profoundly from those of people in another. One group may enjoy singing folk songs, another listening to romantic folktales, and a neighbouring group may even be concerned only with legends and traditions. This difference is often geographical, so that the student of folk literature in the Pacific islands who may later investigate a central African tribe will find a completely different emphasis in the two areas. These differences may well depend upon the varieties of religious concepts held by the group or its natural environment, whether islands or jungle or cultivated farm lands, or its stability or mobility. These characteristics are likely to become especially deep seated in groups that have been settled in one place over a long period of history. Frequently they may correspond to national frontiers, but more often they are aspects of the general culture of an area and may well be quite independent of political or linguistic boundaries.

The Russian epic songs are found only in Russia, but the wonder story such as Cinderella or Snow White is a part of the folk literature of a good portion of the world. The Navaho Indians of the southwestern United States place great emphasis on their remarkable chants and lengthy folktales. Their neighbours throughout the Great Plains tell many well-constructed unified stories but confine their rituals largely to the dance. In Europe the Irish excel in storytelling, both of legends and fictional tales, so that even today it has been possible to record a prodigious number for their national archive. But in England and Wales the folktale is little cultivated, preference having been given to legends and ballads. As expected, there is a contrast between the abundance of oral saints' legends in Spain and Italy and their rarity in Scandinavia. Finland, meeting place of Eastern and Western tradition, shows an abundance of nearly every kind of folklore. From eastern Europe to Central Asia the folk epic flourishes.

Tales and origin legends have been collected in great numbers from various parts of Oceania, where there is a common mythological background extending over enormous distances. Except for probable early contact by way of Indonesia, these folktales seem to show little Eurasian influence. In many parts of South America the merging of Iberian, Indian, and Negro materials seems almost complete.

The folk literature of the North American Negroes is in a state of continual change, reflecting their history. Much certainly goes back to Africa, usually by way of the West Indies, and much was borrowed long ago. But the Negroes have themselves in a truly oral fashion developed songs and stories, and particular music styles. Of very special character is the folklore of modern Israel. Jews coming from various lands to the East and the West have brought together folk literature from all these countries. Assimilation of this is a long task, and since divergent language backgrounds are unimportant for folktales, the problem is to absorb the great variety of forms.

Taken the world over, folk literature is found everywhere, though the emphasis differs from place to place.

MAJOR FORMS OF FOLK LITERATURE

Function of the folk song

Folk song. Some kind of singing is almost universal and it is probable that where there are no reports of it information is simply missing. Folk song implies the use of music, and the musical tradition varies greatly from one area to another. In some places the words of songs are of little importance and seem to be used primarily as support for the music. Frequently there are meaningless monosyllables and much repetition to accompany the voice or the musical instrument. In much of the world drums and rattles, beating time by hands or feet or the stroking of a harp give a strong rhythmic effect to all folksinging. In others, flutelike wind instruments or bowed fiddles of one kind or another affect the nature of the folk-song texts. In many places these apparently meaningless folk songs are of great importance, serving as excitement to war or love or as a part of religious or secular ritual. Through them the group expresses its common emotions or lightens the burden of communal labour. In certain preliterate groups, and sometimes elsewhere, folk songs are used for magic effects, to defeat enemies, to attract lovers, to invoke the favour of the supernatural powers. Sometimes the magic effect of these songs is so greatly valued that actual ownership of songs is maintained and their use carefully guarded. They may come to the owner in a dream or as the result of fasting or other austerities.

Even when folk songs are not used for such practical purposes but only for the pleasure of singing or listening, the greater part of the world uses them for the expression of ideas or emotions held in common by the group. Only in societies used to the songs of composers or poets does purely personal expression enter into the folk song. This is not frequent, and songs of this type are hardly to be distinguished from some of the simple lyrics of poets such as Robert Burns. Folk songs, essentially expressions of commonly shared ideas or feelings, are often trivial but sometimes they may be profoundly moving.

Narrative singing

The lyric folk song in one form or another is found almost everywhere, but this is not true of narrative singing. Unless the reporting of the activities of preliterate cultures has been very faulty, it would seem that the combination of song and story among these peoples has been rare, in spite of a wealth of prose narrative. On the other hand, in major Western and Asian civilizations the narrative song has been important for a long time and has been cultivated by the most skillful singers. In the course of time these songs of warfare, of adventure, or of domestic life have formed local cycles, such as the *byliny* of Russia or the heroic songs of Yugoslavia and Finland or the ballad tradition of western Europe and elsewhere. Each of these cycles has its own characteristics, with its distinctive metrical forms, and its formulas both of events and expression. Any reader of the Homeric poems will be aware of their essentially oral and musical nature, and all the early literary narratives of Sumer and the Near East suggest a long previous development of narrative singing.

Ballad. A special tradition of tales told in song has arisen in Europe since the Middle Ages and has been carried to wherever Europeans have settled. These ballads, in characteristic local metrical forms and frequently with archaic musical modes, are usually concerned with domestic or warlike conflict, with disasters by land or sea, with crime and punishment, with heroes and outlaws, and sometimes, though rarely, with humour. Despite a folk culture fast being overwhelmed by the modern world, these ballads are still being sung and enjoyed.

Folk drama. Belonging only remotely to oral literature is folk drama. Dances, many of them elaborate, with masks portraying animal or human characters, and sometimes containing speeches or songs, are to be found in many parts of the preliterate world. Though the action and the dramatic imitation is always the most prominent part of such performances, these may be part of a ritual and involve speaking or chanting of sacred texts learned and passed on by word of mouth.

The ancient Greek mysteries, as well as secret societies even down to the present, have retained this method of transmitting dramatically their traditions and their teachings and commentary. Some dramatic rituals indeed were not secret but part of a public cult. Thus in ancient Greece the feast of Dionysus led eventually to classical Greek drama, and in the Middle Ages the dramatic celebrations of the Christian Church developed into the medieval folk dramas and at long last into the literary drama of the Renaissance and later.

The medieval mummers' plays and their modern survivals, the Passion plays, the Mexican reenactment of historic scenes such as "The Moors and the Christians," and the modern pageants—all these are based on written texts, however crude, and are beyond the scope of this treatment (see RITES AND CEREMONIES and LITERATURE, THE ART OF).

Fable. Fables, whether of the well-known Aesop cycle, with animals acting according to their real natures, or those from India, where animals simply act as men and women, are literary in origin. But they have had an important influence on folk literature. In addition to appearing in written collections, a number of these are told by storytellers in many parts of the world. Such fables as "The Ant and the Grasshopper," with appropriate morals, have been frequently recorded along with oral tales and have undoubtedly served as models for new animal stories. Sometimes these new tales have eventually received literary treatment, as in the medieval "Reynard the Fox," and then been carried back around the world by storytellers. In such narratives the borderline between folk literature and other literary expression is impossible to draw.

Folktale. The oral fictional tale, from whatever ultimate origin, is practically universal both in time and place. Certain peoples tell very simple stories and others tales of great complexity, but the basic pattern of taleteller and his audience is found everywhere and as far back as can be learned. Differing from legend or tradition, which is usually believed, the oral fictional tale gives the storyteller absolute freedom as to credibility so long as he stays within the limits of local taboos and tells tales that please.

A folktale travels with great ease from one storyteller to another. Since a particular story is characterized by its basic pattern and by narrative motifs rather than by its verbal form, it passes language boundaries without difficulty. The spread of a folktale is determined rather by large culture areas, such as North American Indian, Eurasian, Central and South African, Oceanic, or South American. And with recent increasing human mobility many tales, especially of Eurasian origin, have disregarded even these culture boundaries and have gone with new settlers to other continents.

Variety of types

In many preliterate cultures folktales are hardly to be distinguished from myths, since, especially in tales of tricksters and heroes, they presuppose a background of belief about tribal origins and the relation of men and gods. Conscious fictions, however, enter even into such stories. Animals abound here whether in their natural form or anthropomorphized so that they seem sometimes men and sometimes beasts.- Adventure stories, exaggerations, marvels of all kinds such as other world journeys, and narratives of marriage or sexual adventure, usually between human beings and animals, are common. Much rarer, contrary to the views of earlier students, are explanatory stories. Tales of this description are especially characteristic of Africa, Oceania, and the South American Indians.

In much of the world, especially Europe and Asia, the folktale deals with a greater variety of incidents than just described. In the course of time folktale scholars have given most attention to this area and have classified these stories so that the vast collections of them in manuscripts or books can be referred to with exactness.

All readers of such collections as those of Grimm will easily recall examples of tales of speaking animals. These may be old Aesop fables or parts of the medieval Reynard epic, but most of them are based on some ancient oral tradition. Such animal stories are especially numerous in eastern Europe. But better known perhaps are the ordinary folktales that deal with human beings and their adventures. For these tales, usually laid in a highly imaginative time and place—a never-never land—and filled with unrealistic and often supernatural creatures, there exists no good English word, so that usually scholars use the German term *Märchen.* Here belong "The Dragon Slayer," "The Danced-Out Shoes," the "Swan-Maiden" tales, "Cupid and Psyche," "Snow White," "Cinderella," "Faithful John," "Hansel and Gretel," and their like. Here also belong certain stories with religious or romantic motivation and tales of robbers and thieves—"Peter at the Gate of Heaven," "The Clever Peasant Daughter," "Rhampsinitus."

Jests and anecdotes

A major division of this classification of tales deals with jests and anecdotes. Examples are the many stories of numskulls, of clever rascals, and tall tales filled with exaggerations or lies. Finally come formula tales like "The House that Jack Built."

Among jokes and anecdotes a number are risqué or actually obscene. The indexes of the classification have included only those occurring in the published regional surveys. These surveys, and the books and manuscripts on which they have been based, have been subject to severe editing in order to avoid social or even legal offense. Some of the older anthropologists thought to avoid the eyes of the nonscholar by writing such tales in Latin, but a newer generation is much less squeamish. Folk stories now appear in print covering the gamut of the erotic—tales of seduction, realistic descriptions of normal or abnormal sexual activity, and scatological stories of great indecency.

This index of tale types fits the region for which it was planned and is constantly being improved and expanded, but it was never designed to cover the world. The Eurasian types are usually recognizable in any part of the globe and for them this type index is valuable. But for use with stories on a worldwide basis something less formal is needed, a classification of the possible or likely narrative motifs, minute or extended, and wherever found. Such a motif index has in fact proved useful outside of the Eurasian area, wherever comparative studies are undertaken, for parallels or analogues in simple motifs occur even in far distant places, often presenting extremely puzzling problems.

By use of such indexes and from the labours of many scholars, much material for examination of the folktale is available. These studies have been pursued since the 18th century, though until about 1900 most of them were premature attempts to answer the general question of where folktales come from. Eventually it became clear that no satisfactory solution is available, but that every tale has its own history and can be studied only with laborious attention to detail.

Folktale scholarship

In contrast to a literary story, with its standard text and author living in a definite time and place, the folktale is anonymous. Its originators have long been forgotten and it exists in many versions, all equally valid. Instead of being fixed like a literary document, it is in continual flux. But with hundreds of versions of a particular tale available for study it is possible to establish certain norms of plot structure and to point with some assurance to the varieties of subtypes that give clues to its life history. Such an analytical study of these hundreds of versions usually results in some hypothesis about the original form of the plot and the passage the tale has taken through time and space. In this way some 30 or 40 of the more complicated stories have been studied.

These geographical and historical investigations depend on the fact that the plot of the tale is complex enough to admit of really analytical study. For simpler stories and anecdotes scholars have had to be content with less exact methods, usually resulting in nothing more than accounts of their distribution and the known facts of their history.

Most of the attention of students of folktales during the 20th century has been given to historical questions and to preparing the apparatus for studying them—collecting, with ever improved techniques, arranging and archiving materials from manuscripts or books, and indexing types and motifs, so as to make collections even in remote or difficult idioms available to the serious investigator. But the folktale also has given rise to studies that are not strictly historical.

The attempts during the 19th century to find hidden meanings in tales were generally based upon the theory that they were broken-down myths and had lost their original meanings through linguistic misunderstanding. The result was that this "original meaning" was always found to be some conflict between weather or seasonal phenomena (winter: summer; clouds: sunshine; etc.). This type of interpretation has now generally gone out of fashion and has given place sometimes to explanations based upon ancient rituals or to some variety of psychoanalytic treatment. Though both of these possible sources of folk literature merit examination, the resultant interpretations have usually been merely astonishing to those acquainted with the actual history of the tales studied.

A much more fruitful approach to an investigation of folktales has been the studies of the tellers of stories and their audiences. From these has come an appreciation of the way in which folk literature is carried on in a tradition. A great deal more may be expected from such investigations, usually based on an intimate knowledge of the living lore of a single people.

Structural studies, especially of the folktale, have been engaging the attention of more and more scholars. Though particular plots may occur over large parts of the world, the form and literary style of the narrative is likely to be traditional within certain historical or geographic limits. The direction and strategy of these studies of structure are still unclear, but progress is being made.

Legend and tradition. Generally folktales are considered both by tellers and listeners as purely fictional. The line, however, between belief and unbelief is vague and varies from culture to culture and even from person to person, and even in the most sophisticated societies legends of strange things from the past or present continue being told and are usually believed.

Credibility of the legends

Stories about marvelous creatures are worldwide. Often these are merely mentioned or described and the belief in their existence is taken for granted. Frequently, however, there are circumstantial accounts of meetings with them, which result in adventures pleasant or distressing. With such creatures it is sometimes hard to tell whether we are dealing with a fictional story such as that of the dragon slayer of the typical European fairy tale or with a legend actually believed, such as that of St. George and the dragon. Though the folk in all parts of the world handle these stories with varying degrees of belief, there exists everywhere a remarkable resemblance among these supernatural creatures. The dragon, for example, in something of its characteristic serpent or crocodile form, is of great importance in China as well as in Europe and is represented in both places as a guardian of great treasure. Hardly less well known is the unicorn, and various combinations of man and beast such as the centaur and the minotaur, or the combinations of man and dog, have been a part of the legends of the Old World and occasionally of the New. Giant birds carrying men off in the claws, the phoenix reviving from its own ashes, flying horses carrying men through the air, sirens, mermaids and mermen, and unbelievable creatures resembling these appear in traditions all over the world. There are treasure animals of all kinds, not only the goose that lays the golden egg but the cow that furnishes treasure from its ear. The horse may warn the hero of danger or may determine which of two roads he should take. Important building sites are said to have been determined by the actions of a wise animal. Speaking animals, of course, figure prominently in all folk literature and even in such literary forms as the fable. Animals may speak to each other on Christmas Eve, or they may have governments and elect kings or celebrate weddings. These are only a few of the traditions current with a large part of mankind.

The relation between the animal and the human is very close in all folk literature. In the preliterate cultures of the American Indians, the Pacific Islanders, or the Central Africans, the culture heroes who are responsible for the good and the bad in the life of the tribe may upon one occasion appear as animals and upon another as men. Such was true of the ancient gods of Egypt or Greece. The question whether Coyote of the American Indian tribes is an animal or a man apparently makes no difference to those who tell stories about him.

Supernatural creatures in legends

Aside from these semidivine creatures, now animal or bird or man as they wish, supernatural and ill-defined creatures, much more difficult to visualize, are also common. Fairies or their counterparts appear in the legends of a good part of the world. It is hard to define them, for in one place they will appear in full human size, in another as little creatures inhabiting mounds or caves or living under the roots of trees. In some countries they are benevolent creatures, helpful to men and women. They reward human services but punish misdeeds. They marry or consort with human beings. In some traditions they are malevolent creatures, and meetings with them always bring disaster or bad luck. Almost every country has produced its own variety of helpful and harmful creatures. Stories of the activity of witches and devils, or water spirits and the supernatural guardians of mountains or trees vary in details from land to land, but many of the incidents related about them are easily transferred from one to another. Stories of visits to quite other supernatural realms, fairyland, for example, may be told in all their details in Russia or Greece. Giants are usually considered to be ogres of one kind or another but they may also be considered the most stupid of all beings and may be the subjects of hundreds of numskull anecdotes. Underground creatures like the dwarfs in "Snow White" are usually helpful and kindly, but other underground creatures bring only disaster.

The widespread belief in the return of the dead has resulted in many stories of encounters with ghosts or of actual resurrection. These stories differ greatly in various parts of the world and are much influenced by the current religious ideas. It is likely that in the whole world of traditional literature the belief in ghosts has survived longest.

Historicity of some legends

Traditions of historic characters have a tendency to repeat themselves from land to land and although they are told as facts may form as definite patterns as any fictional folktale. Such stories as Joseph and Potiphar's wife or the exposure and ultimate return of the hero appear in many places. The expected return of King Arthur from Avalon or of Barbarossa from his cavern are only two examples of a widespread motif of this kind.

It is difficult and perhaps impossible to distinguish the explanatory legend from the myth. Tales explaining the origins of customs or of the shape or nature of various animals and plants, of such distant objects as the stars, or even of the world itself often ascribe such origins to the action of some ancient animal or to some magic transformation. These are often connected with stories of the gods or demigods and may even be a part of the religious beliefs of those who tell them.

Generally, legends and traditions of this kind are simple in their form and contain only a single motif or at most two or three. The problem of proper classification for the purpose of studying these has proved very difficult, for while the materials of these legends and traditions show many interesting parallels and resemblances, they vary greatly from place to place. The relation of these stories to actual history, to mythology, and to the fictional folktale is of much interest to students of folk literature.

Proverbs, riddles, and charms. Three of the shorter forms of folk literature—proverbs, riddles, and charms—are not confined to oral expression but have appeared in written literature for a very long time. The proverb that expresses in terse form a statement embodying observations about the nature of life or about wise or unwise conduct may be so much an oral tradition as to serve in some preliterate societies as a sanction for decisions and may even be employed as lawyers employ court precedents. In literature it dominates certain books of the Old Testament and is found even earlier in the writings from Sumer. There has been a continual give and take between oral and written proverbs so that the history of each item demands a special investigation.

Proverbs and riddles in preliterate societies

While the proverb makes a clear and distinct statement, the purpose of the riddle is usually to deceive the listener about its meaning. A description is given and then the answer is demanded as to what has been meant. Among examples in literature are the riddle of the sphinx in Sophocles and the Anglo-Saxon riddles, based on earlier Latin forms. In oral literature the riddle may be part of a contest of wits. But even if the answer is known, the listeners enjoy hearing them over and over. In Western culture the riddle is especially cultivated by children.

Charms, whether for producing magic effects or for divining the future, also exist in folk literature as well as in the well-known Anglo-Saxon written form. The study of these extends over all parts of the world and back to the earliest records.

Children's use of folk literature. As a part of their play activities children not only play old games but repeat counting-out rhymes and retain play-party songs that have long ceased to be a part of adult activity in Western culture. Although the knowledge of those matters is available to children in their books, in actual practice it is passed on by word of mouth or by imitation, and the tradition may spread from school to school over a continent with great rapidity (see LITERATURE, THE ART OF).

STUDY, COLLECTION, AND PRESERVATION

As abundant as folk literature is and has been, its investigation has been seriously undertaken only within the past two or three centuries. The principal difficulty has been the assembling of material on which to base such studies. Its very oral nature makes it impossible for one man to be acquainted at first hand with more than an extremely small part of this activity. It is only when some sort of written record has been made of the oral material that any general studies are possible.

For the still unlettered peoples, the reports of ethnologists and anthropologists, as a part of their general studies of the cultures of widely distributed groups, have often given good accounts of folk literature and have frequently furnished texts of material they heard. Though these reports are extremely uneven and often fragmentary, they do give a sampling of the literary expression of many and diverse parts of the earth.

When attention is shifted to the ancient world before the use of writing, scholars are almost entirely dependent on analogies from the unlettered groups just mentioned. It will never be known what tales were told or what songs were sung by the builders of the Egyptian Pyramids or the temples in Sumer, but it seems fair to assume that even then these peoples were not silent. Of course it must be remembered that they did eventually develop a written literature, so that the analogy with modern unlettered peoples may not be completely valid.

For folk literature since the development of writing, scholars are dependent on several things. There may be specific references in literary documents to the existence of particular tales or songs and often to their manner of production. The Old Testament is a good source for these, and both the *Odyssey* and *Beowulf* contain good pictures of the performances of folk minstrels and bards.

Many collections of folktales and legends, of lyric and heroic songs, and of riddles and proverbs have been recovered directly from popular tradition within the past three or four centuries. When the collection of this material began it was nearly always rewritten in the prevailing literary fashion. Excellent examples of such rewritten tales will be found in the collections of Giambattista Basile (1634–36), Charles Perrault (1697), and various German writers such as Johann Karl August Musäus and Clemens Brentano in the 18th and early 19th centuries. The Brothers Grimm with their *Kinder und Hausmärchen* (1812–15) have as their ideal the exact recording of tales as heard from oral tellers, though it is clear that many stories in their famous work are not folk literature at all. In the same way, collections of folk songs and ballads were severely edited well into the 19th century.

Partly as a result of the Romantic movement in literature and partly of the interest in primitivism and the common folk, the recording of all sorts of songs and oral tales since about 1800 has been phenomenal. More and more the attempt has grown to recover material as it actually exists. Many thousands of volumes are to be found in great libraries that give a good sampling of folk literature in all parts of the world. The last century has also seen the development of large regional or national archives, many of them containing hundreds of thousands of items available for study. All of these books and manuscripts have become increasingly valuable as the techniques of collecting have progressed from casual longhand notes and rewritings through various stages to mechanical recording on discs and tapes.

Preservation of folk material

With mechanical recording it has been possible to assemble properly attested folk literary material from all parts of the world. This improved collecting has proceeded at an impressive pace and makes possible comparative studies of all kinds, based on the oral record.

As for the folk literature of peoples predominantly unlettered, these greatly expanded bases for study have brought out not only the characteristics found everywhere but have pointed up the differences found from place to place. Generalizations formerly accepted have to be reviewed in the light of these differences. With increased collecting, for example, do the likenesses or unlikenesses of American Indian tales and legends become more manifest? Does folk literature in a certain part of the world follow culture areas or language boundaries or some other principle? Such problems can now be investigated with the assurance that modern collectors have made every effort to record the oral tradition as it actually exists.

Much the same may be said of the 20th-century folk literature that exists among literate people side by side with written works. The collecting has improved both in quantity and quality. And not only have libraries been receiving new books of folk literature collections from interested persons everywhere but these collectors are better trained and equipped. The greatest improvement, however, in the study of folk literature transcribed in writing has been the development of folklore archives, of which a large part are concerned with various kinds of oral literature. These are growing rapidly, are scattered over much of the world, and are becoming well indexed and accessible.

(S.T./Ed.)

Folk music

Typically, folk music, like folk literature, lives in oral tradition; it is learned through hearing rather than reading. It is functional in the sense that it is associated with other activities. Primarily rural in origin, it exists in cultures in which there is also an urban, technically more sophisticated musical tradition. Folk music is understood by broad segments of the population, while cultivated or classical music is essentially the art of a small social, economic, or intellectual elite. On the other hand, that widely accepted type of music usually called "popular" depends mainly on the mass media—records, radio, and television—for dissemination, while folk music typically is disseminated within families and restricted social networks. (For further discussion of the differences between popular music and folk music, see POPULAR ARTS.) But the introduction of songs from folklore into the mass media blurs the distinction, and folk music in earlier times may be discussed separately from that of the period after World War II. Moreover, while folk music as defined above exists in all cultures in which there is also a cultivated musical tradition, such as Japan, China, Indonesia, India, and the Middle East, the usefulness of the concept varies from culture to culture. It is most convenient as a designation of a type of music of Europe and the Americas.

ORIGIN, FUNCTIONS, AND TRANSMISSION

Perhaps the most important characteristic of a folk song is its dependence on acceptance by a community—that is, by a village, nation, or family—and its tendency to change as it is passed from one individual to another and performed. This process of cultural exchange is known as "communal recreation."

A piece of folk music is the property of the entire community. But contrary to beliefs promulgated in the 19th century, folk songs are normally created not by groups of people but by individuals. When it is first composed, each song is the work of one composer, though it is recreated constantly by the performers who learn and sing it. The composer may create new songs by drawing together lines, phrases, and musical motifs from extant songs, possibly combined with entirely new ones and with standard opening or closing formulas. In European folk music, a small number of tune types account for most of the repertoire. English folk music, for example, is believed to consist largely of about 40 "tune families," each of which descends from a single song. And the majority of English folk songs are members of only seven such tune families.

There is frequent interchange of tunes between neighbouring countries. A few tune types are found throughout the European culture area. Each country, however, tends to have a repertory of its own, with stylistic features as well as tunes that are not shared with neighbours. Textual types (such as ballad stories) are more widely distributed than tune types.

The 20th century has seen the decline of folk tradition in many areas, particularly those that became heavily urbanized and industrialized. From the Middle Ages until the 19th century, folk music probably had been distributed evenly throughout Europe and the Americas. After 1950, folk music was found most readily in areas that were not heavily industrialized, such as the isolated mountainous regions of North America or of Italy and in the countries of eastern and southern Europe. In the Americas, folk music of European origin became mixed with elements of non-Western music, especially African and (in Latin America) American Indian.

Much folk music can be said to be "functional" in that it is not primarily entertainment or of aesthetic interest but an accompaniment to other activities, particularly ritual, work, and dance. In a traditional folk society, music is a necessity in almost all rituals and festivals. The words of folk song can serve as chronicle, newspaper, and agent of enculturation. In modern industrial nations, folk music is perpetuated by ethnic, occupational, or religious minorities, among whom it is thought to promote self-esteem, self-preservation, and social solidarity. Such functions of folk music have been used by organizations advocating social change, such as the U.S. civil rights and trade unionism movements.

Variants in folk music

Folk music is usually transmitted by word of mouth, or oral tradition. This means that a folk song can change as a result of the creativity of those who perform it or of their particular musical style or of their faulty memory. As it is handed down from generation to generation a folk song develops additional forms, called variants, which may differ markedly from each other. For example, a song with four musical lines (*e.g., ABCD*) may lose two of these lines and take on the form *ABAB*. In turn, two new lines may be substituted for the initial two, giving it a form *EFAB*. Folk tunes also change when they cross ethnic or cultural boundaries. A German variant, for example, may exhibit characteristics of German folk music, while its variant in Czechoslovakia, although recognizably related, will assume the stylistic traits of Czech folk music. The degree to which songs change varies from culture to culture. In some, presumably those that value consistency and object to change, such as western Europe, songs change little and slowly. In others, such as Afro-American cultures, the opposite tendency is found.

In spite of its dependence on oral tradition, folk music tends to be closely related to music in written tradition. Many folk songs originate in written form. For many centuries, popular and classical composers have adapted folk music, and in turn, influenced the oral tradition. A modern analogue of written tradition, recording, substantially influenced the oral tradition, as folk singers could hear various arrangements of folk music in private and commercial recordings. Thus, the transmission of folk music has not been an isolated process but one intertwined with other kinds of musical transmission.

GENERAL FORMAL CHARACTERISTICS

Composition. The composition of folk music differs little from that of popular and classical music, except that most folk songs are composed without notation. The relationships among the sections of folk songs and their scales and rhythms are also found in the other music of the same culture. Systematic improvisation as a method of composition is found only occasionally, as in the epic songs of eastern Europe. It is often difficult to ascertain whether the same composer created both words and music in a folk song, but, in many, they are known to come from different sources.

Among the most important genres of folk music are ballads, generally short narrative songs with repeated lines, epics (longer narratives in heroic style), work songs, love and other lyrical songs, songs of a ceremonial nature accompanying the life cycle of man or the annual agricultural cycle, songs accompanying games, and lullabies. These genres are distinguished usually in their texts, but in some cultures, also in their music. Instrumental folk music is most frequently an accompaniment to dance.

Melodic form. The typical melodic form of European folk music is strophic, that is, a stanza consisting of from two to eight lines (but most typically, four lines) is repeated several times in the song between successive stanzas of the text. The relationship among the lines of the repeated stanzas varies. For example, in English folk music, four lines with different content are common (*ABCD*), but forms whose endings revert to materials presented at the beginning are also common (*e.g., ABBA, AABA, ABCA, ABAB*). Similar forms are found in eastern Europe, where the use of a melodic line at successively higher or lower levels is also important. Thus, in Hungarian folk music, the form AA^5A^5A or AAA_4A_4 (the numbers indicating intervals of transposition) is common. In Czech folk music, AA^5BA is a common form. Despite the variety of arrangement of the musical lines of a song, the textual and musical lines nearly always coincide. In western European folk music, these lines are almost always of equal length; eastern European folk music frequently departs from this principle.

Exceptions to the strophic form

Among the exceptions to the strophic form are children's songs and ditties and epic songs. The former tend to be simple: they use limited scales and rhythms and small melodic range, and they may consist of only one musical line repeated many times. They appear to form an archaic stratum of European music and tend to be similar in musical content throughout the continent.

Epic folk singing is limited to a small number of folk traditions: Balkan, Finnish, and Russian, as well as non-European cultures. The tendency to repeat and vary a musical line many times is also found in epic singing, which is to some extent improvised.

The influence of popular music on folk music, which became very strong in the 19th and 20th centuries, has tended to limit and to standardize forms. The variety of melodic forms is greater, for example, in older English, Anglo-American, German, and Czech folk music than in later music.

Instrumental accompaniment

Most folk music is monophonic (that is, with only one melodic line), but polyphonic folk music, with several melodic lines, is found in some parts of the world. The accompaniment of melody by instruments is widespread as well, though all cultures have many songs that may be sung without it. The accompaniment of folk music in western Europe appears to have changed over the last thousand years. Originally, very simple, perhaps drone-like material was performed by string instruments such as harps, zithers, and psalteries. Later, simple harmonic sequences developed that were closely related to the practices of 18th-century classical music and involved a larger variety of instruments, including guitars, banjos, and string ensembles. The popular folk music of the modern cities embodies still more complex harmonic idioms.

Polyphonic vocal folk music is far more common in eastern and southern Europe than in western Europe. Styles vary from the simple two-voiced structures using drone-like techniques and parallel singing of the same tune at different pitch levels in Italy and the Balkans to the more sophisticated choral songs in three or four voices, found in Russia and Ukraine. Rounds are found throughout Europe. Heterophony—the simultaneous performance of variations of the same tune by two singers or by a singer and his accompanying instruments—is important among the southern Slavic peoples. Parallel singing is perhaps the most common type of folk polyphony: parallel thirds—that is, singing the same tune at an interval of a third—are found in Spain, Germany, Austria, the Czech Republic and Slovakia, and farther east; parallel fourths and fifths are sung in the Slavic countries. Instrumental folk polyphony is geographically more widespread than vocal. Bagpipes, for example, which use the drone principle, are ubiquitous in Europe. Scandinavian vocal music is largely monophonic, but complex styles of instrumental polyphony were developed in the repertoires of various types of fiddles, such as the Norwegian Hardanger fiddle.

It must be borne in mind that certain cultures, such as the British, the Hungarian, and the Cheremis, or Mari people of Russia, while having very little polyphonic folk music, have developed highly complex repertoires of monophonic folk song. Polyphony should not be considered an indication of an advanced state of art.

Rhythm and metre. In folk music, rhythm and metre largely depend on the metre of the poetry. Thus, in western Europe, where poetry is organized in metric feet, there is a tendency toward even isometric structure based on one type of metre—typically, $\frac{4}{4}$, $\frac{3}{4}$, or $\frac{6}{8}$, although $\frac{5}{4}$ also appears. In eastern Europe, generally, the number of syllables per line is the main organizing factor, regardless of the number of stressed syllables. Accordingly, the number of notes but not the number of measures is important, and repeated but complex metric units (*e.g.*, $\frac{7}{8}$, $\frac{11}{8}$, $\frac{13}{8}$, etc.) appear, particularly in Hungarian, Bulgarian, and Romanian songs.

Metre and singing style

Rhythmic structure is closely related to singing style. Singers in the older, ornamented styles frequently depart from rigid metric presentation for melismata (*i.e.*, a single syllable sung to a series of notes) and other expressive effects. Generally speaking, instrumental music is more rigorously metric than vocal. Nonmetric material, some of it consisting of long, melismatic passages, is also found in vocal and instrumental music in parts of Europe influenced by Middle Eastern music, such as the Balkan and Iberian peninsulas.

Scales. Generally speaking, the scales of European folk music fit into the diatonic tone system of European art music. On the whole, the scales of folk music in Asian high cultures are closely related to those consisting of two, three, or four tones, typically using major seconds and minor thirds. These scales are normally used in single-line songs, such as children's ditties, game songs, and lullabies, and they resemble the world's simplest music, that of certain tribal cultures. Among the most important scale types in Europe is the pentatonic, usually consisting of minor thirds and major seconds; it is found throughout the continent but especially in songs and song types not strongly influenced by the art music and popular music of the cities. Diatonic modes are the third important group. The modes most frequently used are Ionian (or major), Dorian, and Mixolydian, but Aeolian, Phrygian, and Lydian are found as well. The Ionian mode is most common in western and central Europe; others are found in eastern Europe, Scandinavia, and England (as well as in English-derived music around the world).

Scales with a predominance of small intervals close to minor seconds are found in the areas once influenced by Middle Eastern music.

INSTRUMENTS AND PERFORMANCE STYLES

Folk music instruments vary in type, design, and origin. They can be divided into roughly four classes.

Four classes of folk instruments

Among the simplest instruments are those that European folk cultures share with many tribal cultures throughout the world. Among them are the following: rattles; flutes, with and without finger holes; the bull-roarer; leaf, grass, and bone whistles; and long wooden trumpets, such as the Swiss alpenhorn. These instruments tend to be associated with children's games, signalling practices and remnants of pre-Christian ritual. They evidently became distributed throughout the world many centuries ago.

A second group consists of instruments that were taken to Europe or the Americas from non-European cultures and often changed. Among them are bagpipes, the folk oboes of the Balkan countries, the banjo, the xylophone, and folk fiddles such as the Bulgarian one-stringed *gusla.*

Another group consists of the instruments developed in the European folk cultures themselves from simple materials. A characteristic example is the *Dolle,* a type of fiddle used in northwestern Germany, made from a wooden shoe. A more sophisticated one is the bowed lyre, once widespread in northern Europe but later confined mainly to Finland.

The fourth group that is of great importance comprises instruments taken from urban musical culture and from the traditions of classical and popular music and sometimes changed substantially. Prominent among these are violin, bass viol, clarinet, and guitar. In a number of cases instruments used in art music during the Middle Ages and later, but eventually abandoned, continued to be used in folk music into the 20th century. Examples include the violins with sympathetic strings found in Scandinavia (related to the viola d'amore) and the hurdy-gurdy, still played in France, and related to the medieval *organistrum.*

The manner of both vocal and instrumental performance of folk music may vary greatly. In general, they differ considerably from Western art music. The sometimes strange, harsh, and tense voice in folk song is no more or less natural—or intentional—than the vocal style of formally trained singers. The manner of singing and the tone colour of instrumental music are among the most important characteristics of folk music. They are less subject to change over a period of time and less subject to influence than other characteristics of music such as scale, rhythm, and harmony.

Speaking very broadly, European folk music is sung in one of two styles, named *parlando-rubato* and *tempo giusto* after studies of east European folk music by the eminent Hungarian composer Béla Bartók. The first style, *parlando-rubato,* is probably older. Stressing the words, it departs frequently from metric and rhythmic patterns and is often highly ornamented. The second style, *tempo giusto,* follows metric patterns more precisely and maintains an even tempo. Both styles are found in many parts of Europe and in European-derived folk music. Using other criteria, the contemporary U.S. folk specialist, Alan Lomax, found three main singing styles in Europe and the Americas. The "Eurasian," found mainly in southern Europe and in parts of Britain, is tense, ornamented, and essentially associated with solo singing. The "Old European," found in central Europe and parts of eastern Europe, is more relaxed. Produced with full voice, it is often associated with group singing in which the voices blend well. The "modern European," found mainly in western European singing of more recent materials, is something of a compromise between the other two styles.

Before the 20th century members of a community probably tended to sing very much in the same style. In the 20th century—probably because of the influence of popular music, radio, and records—folk singers began to develop intensely personalized repertories and ways of performing, as may be seen in the work of popular folk singers.

In the Americas, the influence of African performance practices on Afro-American, as well as other folk music, has been important. Among these are the imaginative use of vocal tone colours, antiphonal and responsorial techniques, and complex rhythmic patterns.

RELATIONSHIP TO OTHER MUSIC

The relationship of folk music to art music became a topic of interest in the late 18th century when Western intellectuals began to glorify folk and peasant life. Folk music came to be venerated as a spontaneous creation of peoples unencumbered by artistic self-consciousness and aesthetic theories and as an embodiment of the common experience of inhabitants of the locale. These traits make folk music a fructifying source for art music, particularly when it is intended to express a particular nation or ethnic group. Another theory is that folk music is not created by the folk but is popular music and art music that has "trickled down" to the folk and undergone various transformations (usually debasements) through oral tradition.

A viewpoint intermediate between these two positions has been widely held since 1950. Folk music is seen neither as merely debased art music nor as an essential component of art music. Rather, it is seen to have a symbiotic relationship to other music in the larger society of which the folk community forms a part. In Europe and the Americas the give-and-take between folk music and art music is well documented. Many folk songs collected in oral tradition have been traced to literary sources, often of considerable antiquity. Folk music has been consciously incorporated into European art music compositions throughout history, especially during periods of "renewal" such as the Renaissance, the late 18th century, and throughout the 19th and early 20th centuries.

Relation to popular music

Folk music is closely related to popular music in several ways. Societies possessing popular music also have a folk music tradition—or remnants thereof. The partial duplication of repertories and style indicates such cross-fertilization that a given song may sometimes be called either "folk" or "popular." With reference to music, the terms folk and popular are two points on a musical continuum, rather than discrete bodies of music. From a sociological viewpoint, however, folk and popular music have less in common. Unlike folk music, popular music is primarily produced by professionals for consumption by an urban, nonparticipating mass audience. The vital criteria of folk music (*i.e.*, oral tradition, communal recreation, etc.,) are not operative.

Folk and popular music tended to merge in the 20th century. As folk societies came increasingly within the purview of modern urban society, oral tradition was supplemented or supplanted by the radio and phonograph record. Some folk music thus transmitted maintains stylistic authenticity, but some assumes the characteristics of popular music. Much of what is called folk music in English-speaking countries is a subcategory within popular music. It is the product of urban professionals who appropriate authentic folk music styles for concert and recorded performances.

Relation to jazz and rock music

There has been some interaction between folk music and rock music, as the generic designation "folk-rock" indicates. Folk-rock arose in North America in the 1960s. In its texts, it is modern urban folk song, with topical subject matter, often on social and moral issues. Musically, however, it has the characteristics of rock in its electrified string band and percussion accompaniment. Other current music that mixes folk and popular elements includes: African high life, American jazz, rhythm and blues, country-western music, and many Latin American forms, such as the tango and bossa nova.

Relationships between church music and folk music must also be noted. Some church music derives from the application of religious texts to secular folk tunes. This practice may be seen, for example, in the hymns of the Protestant Reformation and in the revival hymns of 19th century American camp meetings, which were called "folk hymns" because of their origins and associations with folk-like groups.

There are many types of folk dance, some widespread throughout Europe, others peculiar to nations and regions, each with its typical musical style. Certain musical forms appear most typically in the folk dance music of various parts of Europe. Most prominent is a form type in which each line is repeated once, with a minor variation, usually at the end—*e.g.*, $A^1A^2B^1B^2C^1C^2$. Vocal dance music also exists, and in northern Europe even narrative ballads were used for dancing.

STUDY AND EVALUATION

Knowledge of the history and development of folk music is largely conjectural. Musical notations of folk songs and descriptions of folk music culture are occasionally encountered in historical records. Such records, however, show not so much the history of folk music as the history of ideas held by the literate classes about folk music. It is assumed that throughout history literate society has possessed a musical culture different from that of their unlettered contemporaries. Their reaction to folk music frequently was one of indifference and, occasionally, derision and hostility. In medieval Europe, under the expansion of Christianity, attempts were made to suppress folk music because of its association with heathen rites and customs. Uncultivated singing styles were denigrated; Thomas Aquinas expressed a common sentiment when he likened artless singers to beasts. Some aspects of European folk music, however, became assimilated into medieval Christian liturgical music, and vice versa.

History and development of folk music

During the late 15th and 16th centuries, the literate urban classes responded more favourably to folk music than they had in the medieval period. The humanistic attitudes of the Renaissance, such as the elevation of nature and antiquity, encouraged the acceptance of folk music as a genre of rustic antique song. Some music in Renaissance manuscripts is presumed to be folk song by virtue of its musical simplicity and the rural and archaic evocations of its texts. It may, of course, have incurred stylization and change. Renaissance composers made extensive use of folk and popular music. Typical genres include polyphonic folk song settings and folk song quodlibets, or combinations of familiar songs. Folk tunes were often used as structural and motivic raw material for motets and masses, and Protestant Reformation music borrowed from folk music.

Folk music seems to have receded somewhat from the consciousness of the literate classes during the Baroque period. Folk song material in the music manuscripts and prints of the priod is scarce, and there is less folk influence in cultivated music, with the notable exception of stylized dance-music forms.

During the late 18th century folk music again became important to art music, especially among the Viennese classicists. They incorporated folk tunes and the general style of folk music into their instrumental music. The growth of national historical consciousness and the idealization of the rural milieu led to the collection, preservation, and study of folk song in the late 18th and early 19th centuries. Folk song came to be considered a "national treasure" and of considerable artistic merit vis-à-vis cultivated poetry and song. National and regional folk song collections were published. Revitalization of folk music became a means of promoting nationalistic sentiment and a conservative ideology. Governmental encouragement of folk music became common after the early 19th century.

Scholarship. The search for origins and processes of development that motivated much 19th and early 20th century intellectual activity was reflected in folk music scholarship. Among the influences on research in folk music in the 19th century were anthropological concepts of cultural processes and the theory of evolution. Many scholars believed folk music to be a repository of archaisms—a legacy from which the prehistory of music, language, literature, and other cultural traits could be adduced. While later scholars concede that some traits of folk music may be centuries or even millennia old, they are less inclined to speculate on the age of archaic elements of folk music or to offer historical reconstructions other than tracing variants of individual songs or types of songs.

Scholars who specialize in folk music usually have training in ethnomusicology, a discipline concerned with elucidating music in a cross-cultural perspective. Research in the words of folk song remains the province primarily of folklorists and students of language and literature. Folk music theories are concerned mainly with how folk genres and styles and individual folk songs originated, and how, if, and why they change when diffused. Theories of folk music have been beclouded by the difficulties in recognizing, isolating, and defining a phenomenon as elusive and complex as folk music.

Contemporary scholarship

Since the last decade of the 19th century, folk music has been collected and preserved by mechanical recordings. The application of print and recording technology to folk music has promoted wide interest, making possible the revival of folk music where traditional folk life and folklore are moribund. Folk songs are frequently a part of public school music curricula; various clubs, organizations, and societies focussing in one way or another on folk music, often in conjunction with folk dance, have arisen; festivals of folk music and dance are an annual event in many communities throughout the world; conservatories of music have been established for the training of folk musicians, particularly in the Socialist nations; radio stations devote substantial portions of their programming time to the broadcasting of folk or folklike music—again, particularly in Socialist nations.

The literature on folk music is sparse in theoretical works, in historical studies, and in materials integrating and comparing the various styles of folk music in Western culture. There is a great deal of literature showing the relationship between folk music and cultivated music. Most plentiful, however, are collections of music and texts, particularly of individual countries or regions, and even of individual singers. These collections are useful for scholarly comparisons of melodies; they give an imperfect picture of performance practice, however, because Western notation cannot give a detailed description of all aspects of music. After World War II, the availability of commercial records did much to fill this gap, and archives of field recordings were developed at many institutions throughout the world. In the U.S., those of the Library of Congress and Indiana University are most important; national archives exist in most European countries, and particularly in Hungary, Czechoslovakia, Germany, and Scandinavia. Such archives provide ample research material for an enormous diversity of projects. Research has usually dealt with "authentic" (*i.e.,* older) material not heavily influenced by urban popular music and the mass media. Popular folk music has not been studied widely. Several organizations for the study of folk music exist, particularly the International Folk Music Council and the Society for Ethnomusicology. (B.N.)

Folk dance

Although the term folk dance is most commonly applied to the gay, recreational dances of various nationalities, its precise meaning is the subject of much debate among scholars and has not been fully resolved.

Furthermore, scholars and dancers differ in what they admit under the label of "folk dance." One may see folk dance as the traditional dances of a country that evolve spontaneously from the everyday activities and experiences of its people. Another may define it as embracing only dances with magical and economic functions, or as comprising all nonprofessional dances.

The discussions dwell upon the confusion between such terms and concepts as "folk dance," "primitive dance," "ethnic dance," and "stage dance" and on the distinction between folk dance and modern recreational forms of ballroom dancing. (This subject is also addressed in the article POPULAR ARTS.)

Remnants of primitive dance persist in Africa, Oceania, and South America, among peoples who have retained some degree of their traditional religion and ways of life. Such dance throws light on the origins of dance of the Western world. In its retention of its original functions, primitive dance is distinct from the dances of more developed cultures, which may fluctuate between ritualistic and recreational purposes.

Varieties of folk dance

The term ethnic dance seems flexible. Some authorities see no difference between the terms ethnic dance and folk dance. The eminent American dancer Ted Shawn, however, would have ethnic dance subsume folk dance as a subspecies. He considers pure, authentic and traditional racial, national, and folk dance to be "ethnic"; he calls the theatrical handling of them "ethnologic," and he refers to the free use of these sources of creative raw material as "ethnological." Although these distinctions are not hard and fast, they reflect the trend of much ethnic dance toward professionalization. In still another view, folk dance is the dance from which the art dance of a nation inevitably grows, both in technique and in spirit. This concept is particularly applicable to such nations and regions as Japan, India, and Andalusia, where art forms of the dance were a natural outgrowth of the traditional dances.

Purists are disturbed by a trend toward the deliberate "staging" of folk dances, and especially by their increasing professionalization: they might call the adaptations folkloric. Professional dance and secular folk dance have been distinguished as one might separate art from craft, even when the scenarios and choreography of modern dance and ballet adopt materials from folk dance or the larger field of folk culture. Most scholars, however, exclude from folk dance the dances of the commercial theatre, television, and film. Though they generally consider jazz dancing an American folk style, they would exclude formal choreographies in jazz style.

These selected points of view indicate the fluctuating boundaries of folk dance, especially in reference to its functions. Although patterns and movement styles are significant, the function and locale of folk dances have greatest weight in distinguishing them from primitive and theatrical manifestations. Frequently the dances of rural peoples reveal their ritual origins on certain occasions, though they also serve recreational purposes. The origins may be very ancient. Generally, but not always, dances favoured in urban centres have secular purposes and may be of recent, perhaps consciously creative, origin. As in the case of folk song, the origin need not be anonymous, though usually it has been lost in the passage of time. Folk dances have grown out of creative inspiration, and they continue to sprout from the imaginations of individuals and groups, people of all classes who sense the traditions and the aspirations of their environment.

NATURE AND FUNCTIONS OF FOLK DANCE

Functions. Many folk dances best reveal their ancient functions when performed in their native habitat. Outside this context, in a school gymnasium or on a stage, they lose their aura, but on the village green the British Morris dances and the Abbots Bromley Horn dance speak of renewed May Day vegetation and of Paleolithic elk worship. Again, some dances serve various functions. The Spanish Aragonese jota is best known as a rural entertainment for men and women, but it may enliven funerals or appear on American stages.

Dance as ritual and entertainment

The above British examples reflect the transition from pagan to Christian religions and, in more recent times, the change from the attitudes of village and agriculture to those of town and industry and the consequent changes in social relations. As the English scholar Douglas Kennedy pointed out, when primitive religion weakens, some of the

Fusion of function and art form

mystery and the magic departs from the dances that express it. The dancer becomes less a medicine maker than a performing artist as ritual changes imperceptibly into art. In short, man's social adjustment to the environment, for purposes of survival, created both the original dance rituals and their subsequent functional or formal changes. Vestigial animal dances echo ancient animistic rites. The Ainu tribes of northern Japan still mime bear and fox hunts, portraying the animals very realistically. In West Africa, an antelope hunt in dance has ritualistic overtones, while monkey mimes are for entertainment alone.

The Balkans and Central America represent a far-reaching example of adjustment and change. These far-removed parts of the world share ecological circumstances, notably a basically agricultural civilization. Geographically, both narrow into bottlenecks connecting two continents; both combine high and rocky mountain ranges with agricultural lowlands and uplands; both bulge into peninsulas rich in culture. Both have submerged their ancient religious customs to innovations, those of Roman Catholicism and, in the Balkans, of Islām as well. Yet both have maintained their ancient native customs with such compromises as those to the events of the Christian calendar, Christian names, or Islāmic styles. Recently both areas have been receptive to the influx of 19th-century secular European dance forms and have transmuted these importations to suit the native styles.

Combat and agricultural dances. In both areas three dance types show varying degrees of modernization. One type, which takes the form of combat, remains highly ritualistic, albeit with a mixture of pagan and Christian elements. A second, agricultural in function, involves more of the community than the combative type and fluctuates between celebrations of sowing and harvest and of social festivities. The third type, derived from central and western Europe, is completely secular and social.

Male combat dances of the Balkans echo ancient pre-Christian rites for initiation into brotherhoods, the heralding of spring and of animal fecundity, and healing. Fierce battles ensue at the seasonal rituals of the Macedonians, of the Slovenes, and of the Romanians. Animal maskers and buffoons enact resurrection dramas. Along the coasts of Croatia and Dalmatia the battling factions have, under Christian influence, been renamed Moors and Christians or Moors and Turks. These battle dances have related forms and styles in other European countries, from Spain to Great Britain. They also have relatives in Central America, where early Spanish missionaries introduced Moors and Christians to replace the earlier ritual combats of the Indian populations.

Rural celebrations of planting and of harvests feature communal round dances, such as the kolo of the eastern Balkan region, the *horo* of Bulgaria, the hora of Romania, and a variety of Greek chain dances. The celebrations include vestiges of ancient vegetation festivals, impersonations of fertility deities, and "rain magic." The same rounds, however, appeared also at weddings and other secular or semisecular gatherings. Such rounds survive in the mountains of Mexico as *mitotes.* Although they concluded most Aztec and Mayan ceremonies, they have become scarce since the Spanish Conquest. They are still performed to procure rain and an abundant harvest.

Polka- and waltz-based forms

Secular forms. Rural and urban gatherings include the social square dances and couple dances for men and women. Within the last century the Bohemian polka, the Austrian waltz, the Polish mazurka, and the Hungarian czardas have appeared in the northern Balkans. In Central America similar social and courtship dances have become increasingly popular. Each region has a version of the dances known as *jarabe* or huapango. The *jarabe tapatío* of Jalisco, better known as the Mexican hat dance, combines steps from many European nations. All regional dances use polka or waltz steps and European music. With their lively and showy styles, these couple dances are suited to stage performances and occur as such.

In other parts of the world, folk dancers are shifting from a man–deity and man–nature purpose to a man–man or male–female attitude. This is noticeable not only in adaptations of former dances of supplication but also in dances miming agricultural and other occupations, as the Polish sowing of rye and oats, the Hungarian haymaking, the Swedish flax reaping, the clothes washing of Denmark, and the spinning mime of Spain. Some of these occupational dances derive from enactments by medieval guilds or from the mime in medieval branles. They survive as entertainment in adult couple dances and in children's games, often in settings remote from their origin.

In the course of centuries, changes in the beliefs and in the methods of producing the essentials of life have produced numerous adjustments such as the adaptation of the calendar from a basis in agricultural ecology to a basis in Christian festivals and the resultant shifts in the organization of dance groups.

Occasions for dancing. Notwithstanding the trend toward sociable and theatrical objectives, many folk dances celebrate original festivals. In Europe and Europeanized America, however, they show many adjustments to the Christian feasts. In the Balkans, Austria, and other countries the long series of dances for renewed vegetation and life now celebrate Epiphany (Twelfth Night), Carnival, Easter, Pentecost, Corpus Christi, and St. John's Day (June 24). As noted previously, the midwinter dances emphasize male combat and animal impersonations, whereas the springtime dances dwell on new vegetation, in southerly climates on first fruits. Two festivals are particularly spectacular—Carnival and Pentecost.

Festival dancing

Carnival festivals of Europe and the Americas precede Lent, filling the three days before Ash Wednesday. In Austria they perpetuate many pagan dances, particularly in Innsbruck and Imst, with the masked and ghostly phantoms and witches and noisy processions with songs, bull-roarers, drums, and whips. In Spanish and Latin American villages the unruly characters enact a more orderly "combat of winter and summer," in the guise of the ancient Moors and Christians, with the obvious victory of summer. Devils and deaths (*diablos y muertes*) are also on the loose in the role of buffoons. Morality plays are relics of medieval ideology, with speeches in the local vernacular and decorous steppings of Sin, Death, the Devil, Pastorcitas (shepherdesses in white communion dress), and masked animals from the Garden of Eden or bears or tigers.

Urban carnivals bring out animal maskers, deaths, and devils, without ritual connotations in, for instance, Munich. The famous carnivals of Rio de Janeiro and New Orleans draw huge crowds of tourists to observe the masking, competitive parades of floats, and street and ballroom dancing. In the Brazilian medley the street and ballroom dances show interesting contrasts: the samba in the streets is ecstatic, improvisatory, and disorderly, whereas the samba of the ballrooms is more sedate and has set steps. Such urban carnivals have lost sight of the original ritual purpose.

On the other hand the observances of Pentecost, the springtime feast that falls 50 days after the Christian Easter, fit the dances into a framework that meaningfully combines Christian and pre-Christian, New and Old Testament, forms. The Jewish Shavuot festival follows by the same period the Passover, which often coincides with Easter. The Pentecost, known also as Whitsunday, has since AD 200 commemorated the descent of the Holy Spirit on the Apostles, and the Shavuot, originally a feast of thanksgiving for first fruits, has been associated by rabbis with the giving of the Law at Sinai. Both express the joyous resurgence of animal and spiritual powers and of new vegetation.

Harvest dancing

In the southerly climates the festival may already celebrate the first fruits. Everywhere Jewish celebrants bring offerings of fruits and flowers to the temple, with chanting and prayers. In Haifa, Israel, white-clad youths and maidens dance and sing. In the Balkans girls dance for Pentecost, and the community winds in snakelike kolos. In England the community circles around a tree, then around the church, or it holds a maypole dance. In some villages, such as Bampton-on-the-Bush and those of the Cotswolds region, "Morris men" dressed in clean white caper and leap in a procession or in double files, waving white kerchiefs or green branches. The dancers may have the company of clowns, a Jack-in-the-Green clad in greenery. In some English villages and in British-inspired American locations, such dances take place on May Day rather than Pentecost.

Agricultural festivals, especially harvests, may adjust their dates not only to the local climate but to the particular year's weather. The Iroquois Indians of New York State and Ontario adjust their calendar to the ripening of the crops of berries, beans, and corn. They may hold their thanksgiving rounds for green corn between the third week of August and the middle of September. The square dances of the American farmers were held on the occasion of husking bees—before combines took over the work—whenever the corn was ready. Farmers continue their square dances, or "country dances," in barns or in grange halls at odd times or even weekly. Their urban imitators perpetuate these dances assiduously when square-dance and folk-dance societies, often mingling the traditional American dances with those of immigrant peoples, meet in national halls or centres, school or college gymnasiums, or other locations. The gatherings of these enthusiasts and analogous groups on both sides of the Atlantic are legion.

Certain secular or semisecular celebrations adhere to a definite date. Such political holidays as the French Independence Day (July 14) and the Mexican national holiday (May 5) and Independence Day (September 16) feature regional dances outdoors and at indoor balls. The Guelaguetza at Cerro Fortín, Oaxaca, formerly a ritual festival, now combines religious and regional dances for the general public on July 16. Such festivals attract vast numbers of dance teams, native visitors, and tourists.

FORMS AND TECHNIQUES

Social roles in folk dancing

Organization of participants. Although attendance at such public fiestas is haphazard, the participants in many dance gatherings observe closely knit organization and definite rules for the individual's place in the community and in the communal dances. The men in European combat dances belong to a sworn brotherhood of ancient origin. The male and female members of a Mexican votive society, the Concheros, have an intertribal hierarchy paralleling that of the forces of the conquistador Cortés, headed by a *capitán general.* In second rank are the officials of each local group, first and second captains, sergeants, standard bearers, each with specific duties, followed by the common rank of *soldados* and, finally, such attendant characters as Cortés' interpreter-mistress Malinche, the devil, sorcerers, and mythological figures. At present they do not regulate their rituals according to the calendar, though their ancestors probably did.

Although such societies cut across family ties, other organizations are based on descent, especially among American Indians. Iroquois and Pueblo Indians group their clans into two moieties, or halves, of the entire social scheme, matriarchal and patriarchal respectively. In their ceremonies and social events the Iroquois stress the interaction of moieties, with the alternation of moieties in the dance file. However, the New Mexican Pueblos usually feature separate dances for the two moieties and even assign festivals of the two seasons to the summer and winter moieties.

Sex and age roles. These same tribal groups also observe strict regulations according to sex. Iroquois women manage the summer rites for agriculture; the men manage fall and winter ceremonies for animals and cures. Among the Iroquois as well as the Pueblo, men and women hold esoteric dances separately, or men occupy one-half of the dance line and women follow in the second half. In less sacred dances and always in social rounds men and women alternate. Observers report similar customs not only among the natives of the New World but also in the Old, as in Serbia and Great Britain. Men perform the traditional Morris and sword dances, but the sexes mingle in country dances, reels, and quadrilles. The solos in Scottish sword dances are traditionally male performances, but, as a nonauthentic deviation, girls may now execute the tricky steps of the dances.

The traditions of age grades are also becoming diluted. From Greece to New Mexico, almost universally, the older, experienced men and women are the leaders, while the children bring up the rear of dance lines as apprentices. Warrior societies of Great Plains tribes of the United States once observed strict gradations of dance rituals according to age. But these societies are all but extinct, and public war dances admit all ages of males, with females in the background. With the dissemination of folk dances into the schools, children are learning adult routines. However, in remote villages of Europe youngsters have their special dances, and adolescents may enter the adult circles modestly.

Individual and ensemble dances. Generally the individual is submerged in the larger society and must fit into the dance group harmoniously. Some peoples, the Pueblo Indians for example, uphold strict standards of restraint, and within the natural variations of greater or less energy, a member of a dance group should not show off. However, other peoples such as the Iroquois appreciate improvisatory clownery or virtuoso display by talented males. In the Balkans the male leader of a dance line may engage in acrobatics—crouches, leaps, or pivots—while the rest of the group adheres to thè traditional steps. In the Basque provinces of Spain, in the Ukraine, and in Poland male experts have the opportunity to display high kicks or spectacular leaps. The improvisations of these privileged experts have often led to the introduction of permanent new elements into the dances.

Styles of movement. Function, sex, and age all have an effect on a dancer's style of movement. Other psychological factors of group and individual temperament and mood have, for untold centuries, determined the quality and the type of steps and gestures. The climate and topography may have had an effect on the development of regional styles.

Carriage of the body

National and regional style. According to Douglas Kennedy, the ideal of English folk dancers is to hold the body in a straight line from head to toes, creating a vertical equilibrium that makes the dancer light on his feet. This uplifted carriage allows him to reach out and form the contact essential for a unified dance ensemble. This ideal would apply to many folk dance types of Europe, the United States, and Canada, and to some Asiatic round dances, but it does not fit the more dramatic dances of the British Isles nor myriad dances in other parts of the world. Even within England, Kennedy points out the frequently bent-up position and the power of male sword dancers.

In Spain the erect ease of Aragonese line dances contrasts with the swaybacked incisiveness of Andalusian flamenco dances. In Yugoslavia the eastern villages have acquired the vibrations of Turkish dancers, and gypsies use more undulating movements than the Serbs about them. In Asia such hill tribes as India's Todas circle with simple steps and an erect posture, whereas demon dancers of the Pariah caste stamp and leap, and the practitioners of the ancient *Nāṭya* style combine elaborate, symbolic hand gestures with body sways and stamps.

Although India's caste system has produced extreme contrasts, differences in occupation and social class have everywhere affected the spirit and quality of movement. During the late Middle Ages and Renaissance the courtiers who borrowed such rural dances as the branle and the bourrée watered down their rustic vigour. In 19th-century colonial California the descendants of upper class Spaniards performed the polkas, mazurkas, and waltzes with an elegance that contrasted with the rowdy renderings by the gold miners.

Variations by sex. In modern square dancing the difference between male and female styles is negligible, but in most folk dances the women move more gently than the men, with smaller steps, lower leaps, and less raising of the knees or feet. The women dancers have a more sinuous, alluring style in southern Spain; they spin gently in the Austrian and Bavarian *Schuhplattler* and the Caucasian *lezginka,* while the men jump, clap, and shout. Among American Indian tribes the women have a more subdued style and often special, tiny steps except in couple dances that have been adapted from the mainstream of Western social dancing.

Ecological influences. The setting affects the movement style. Joan Lawson suggests differences due to the natural

Influence of the physical environment

environment—a theory that will need more investigation. She maintains that in rich agricultural plains or river valleys, such as the Danubian Plains and parts of France, and Denmark, movements are accented downward as if the body were being drawn toward the soil. Dancers perform in large groups, using the same step, closely linked together by fingers, hands, elbows, or shoulders. By contrast, in mountainous areas there is a good deal of leaping and individual display, especially among the males.

Mimetic and abstract movements. Regional variations include preferences for mime or for abstract movements. India's folk dancers and, half a world away, those of Scandinavia favour mimetic gestures, respectively graceful and comic. Serbia's peasants are interested in purely decorative steps and Ireland's experts are fond of tricky solo steps or complex group patterns that are in no way imitative of outside phenomena. In general, the mime of folk dancers is stylized, having lost the realism of the primitive animal impersonators and of actors in folk dramas.

Opportunities for mimetic dancing are drastically reduced when the hands are required for other formal patterns of the dances. Most folk dancers use their hands and arms for contact in circles, lines, or couples; they wave kerchiefs, as along South America's Pacific coast; they swing soft balls in complex patterns, as in the *poi* dance of the New Zealand Maoris; the women swirl full skirts, as in Spain and Mexico; or everyone lets the arms hang loose or places hands on hips, thus emphasizing foot and ground patterns.

Stylized gestures and motions

In India, dance-dramas based on the life of the god Krishna (Kṛṣṇa) are enacted in Manipur by young women who use simplified gestures descended from the large, complex system of hand gestures known as mudras. The basic gestural symbols derive from the wrist position, the position of the palm, and the poses of the fingers. Each gesture has its prescribed musical accompaniment. A trembling leaf, for example, is symbolized by the *alapallava,* a rotation of the wrist accompanied by a folding and unfolding of the fingers. In Hawaii, a few older women can execute hula gestures clearly descended from the mudras, but younger dancers have diluted the tradition by introducing purely decorative gesture.

In Scandinavian countries male and female imitators of occupations likewise stylize their harvest motions. The youths who portray rough-and-tumble fights, as in the Swedish oxen dance, duel good-naturedly, pull each other's hair, and pretend to box one another's ears. In this last gesture, as in the German *Watschenplattler,* the aggressor merely pretends to touch his opponent, who claps his hands to simulate the blow.

Slavic men and some other skilled performers use steps recalling former animal mime, as the goatlike caper or cabriole, the pawing horse-step or pas-de-cheval, the side-kicking, cowlike rue-de-vache, and the feline pas-de-chat leap. But folk dancers of many nationalities exploit the imageless mazurka or variants of the polka, waltz, and twostep, all in appropriate rhythms. The walking, running, sliding, skipping, or jumping movements are so universal in folk dance that they cannot, by themselves, be considered mimetic.

On the one hand, line dancers of a single region may develop intricate variations of a basic step. Lawson identifies 15 ways of performing the basic kolo step, a step-to-the-side and close. The variants include gliding, swinging of the free leg, crossing, jumping. On the other hand, a widely disseminated step may appear in many forms in different regions. The triple-time waltz is step-together-step in Austria, with pivots at specific times. As the Mexican *atole* step it is forward-back-forward; in the Venezuelan joropo every first beat is heavily accented. As a ballroom dance it reveals diverse patterns: as a propelling step in Spanish and New Mexican quadrilles, a light-footed waltz may balance from side to side, progress forward or backward, or go round and round.

The type of step depends also on the purpose of the dance, whether a solemn processional or an exhibition of skill in leaps or crouches; on the sex or age of the various participants; and on the type of ground plan.

Ground plans. Simple circling leaves the dancer's attention free for elaborate steps, whereas complex ground plans take the mind away from stepping and necessitate the simplest kind of progression by walking or running. Throughout the world the erstwhile ritual dances may involve a simple run, as in the Iroquois corn and bean dances and the serpentine stomp that spread from the ancient Aztecs to Indian agriculturalists of the U.S. Choreographies may combine complexities of step, of rhythm, and of ground plan, like the "game animal dances" along the Rio Grande, but as a rule they emphasize one or another factor.

Serpentine and meander dance formations

The Balkan chain dances feature intricate steps and rhythms, but simple formation of closed or open circles. During closed rounds the men and women remain within the same spot as they inch along counterclockwise. Likewise, participants in French branles circle on location, usually clockwise—the typical direction of northwest Europe. In chain dances the circle is not closed. A leader guides the line, linked by hands or a prop, in meanders and spirals perhaps across open fields. On reversal a tail man will guide the meanders. Such serpentines, of ancient origin, are favourites in the Near East; throughout Europe, especially as the French farandole and the Catalan sardana; in North America, in both native and Europeanized dances; and in such parts of Asia as Manipur. They predominate among agricultural peoples, for they originated in chthonic symbolism.

A specialized form of meander is called the hey in England. Two lines of dancers weave past each other in opposite directions. In a circular fomation this is known as the square dance "Paul Jones" or, if the participants are attached by ribbons to a central pole, as a maypole dance. Here the two opposing groups are or should be male and female. The most elaborate form akin to the hey is the *kolattam,* a stick dance of South India. In the *pinnal kolattam* the dancers weave in and out, at the same time striking short sticks in precise patterns. (The intricacies were diagrammed by Hildegard L. Spreen: see Bibliography.)

Linear dance formations

Dances in two parallel lines have a more limited distribution. As in the case of rounds, the performers may start shoulder-to-shoulder or aligned in the same direction. The lines may cross over or circulate in opposite directions, or pairs of dancers can cross directly or diagonally. Morris dancers use a large vocabulary of interlacings, which resemble those of the American "Virginia Reel"; respectively, the participants are men only and men-and-women. Multiple parallel lines of men and women are customary in Southeast Asia and the South Pacific; Cambodian girls display elegant poses; Balinese men carrying spears mass together in the *baris* dance; and while executing the warlike gestures of the *peruperu,* Maori men remain in one spot.

As noted previously, ritual principles often dictate that in more sacred dances the sexes be separated, whereas in more secular dances they usually alternate or are aligned face-to-face. In modern folk dances, couples circulate within circular formations, as in Moravian rounds and American square dances, or in the extremely elaborate Irish reels of eight couples. In ballroom dances couples generally ignore any geometric designs, and individuals ignore the rest of the group.

In the "possession rite" of Ghana's Akhan society, circle dances by devotees, frenzy dances, and circling by everyone alternate with prayers, chants, offerings, and speeches. A similar structure is evident in the possession dances of Brazil and Trinidad and of the Christian Holiness services in the U.S.

Couples and solos. In the course of history the general trend, during secularization, has been toward increasing complexity, from round or double file to quadrilles, and then from cohesion to a breakup into couples and solos. This disintegration is distinct from the individualism that may be present in primitive dances, for there the soloist had a mimetically compulsive, even priestly, function and was the focus of group activity. Concurrent with the elaboration of patterns, the symbolism has been disintegrating. The vegetation symbols of meanders and arches have been lost, but the designs remain. Face-to-face formations and couple arrangements retain meaning as courtship actions, and despite the loss of the modern folk dancer's relation to, or attempt to act upon, the physical environment, the social contacts between dancers remain.

The type of ground plan affects the contacts between not only the dancers but also the dancers and the spectators. Square dances offer the maximum possibilities of intermingling within a formation, but they exclude spectators. Chain dances lack the give-and-take, but they may wind about or through the spectators, who may enter at any time. Contact, whatever form it may take, is essential to folk dance.

ACCOMPANIMENT TO THE DANCE

The evolutionary process in the relations between the dance and other arts is very similar to the development of the dance itself. From the nearly total integration of dance and life in primitive ritual to modern rock-and-roll, many factors—the passage of centuries, the change from animism to Christianity, the shift from hunting, agriculture, and handicraft to industrialization, the trend from country to city, from sanctuary to village green to stage—have exerted a profound influence on the totality of dance experience.

In the esoteric dance rituals of Australia, in the mythological dance enactments of India and Indonesia, in Nigeria and such of its distant New World derivatives as the vodun, dance is immersed in the larger drama of the rite. The symbolism of the movement patterns is locked into the symbolism of song texts, the traditional music, and the meaning of masks and costumes, not to speak of the setting in a sacred grove. Here and there the decorative invocations to animistic spirits have survived, mysteriously, in the masked animal ghosts of the Austrian Alps, as well as in the "game animal dances" of New Mexico's Tewa Indians. Perhaps these vestiges are not really folk dances. Perhaps folk dances—that is, dances of the people—do not require the integration of all of the arts for gatherings or programs.

Music. In general, the musical accompaniment to folk dances has persevered fairly well. In village and urban hall the devotees use the tunes intended for particular routines, though these tunes may be played on modern instruments. Morris dancers usually preserve the traditional order of a suite—Laudnum Bunches, Bean Setting, Rigs o'Marlow, Shepherd's Hey, Constant Billy. In the execution of isolated kolos, ländler, or country dances, natives and imitators fit the steps to traditional tunes, to live music, piano, or recordings, which may feature old-time clarinets, tabours, drums, and even band arrangements or accordions.

The coordination of tempo and rhythm between dance and music is rarely problematic. It is easy to follow the slow and the fast tempo of a set like the Norwegian *gangar* and *springar,* or the acceleration of an Israeli hora. It is easy to follow the metres of the polka, of the waltz with its accent on the first beat, or of the mazurka with its accent on the second, although the melody may have independent rhythms. It takes more skill to follow some of the Bavarian tunes that shift their metres, and it takes an expert to follow the unusual metres of Greek and Serbian dances, especially when the phrases of the tunes overlap the phrases of steps.

Self-accompaniment. It takes practice also to provide self-accompaniment in rhythm or melody. Rarely do folk dancers provide their entire self-accompaniment, as do the Mexican *viejitos* who play small stringed jaranas, or Hawaiian hula dancers who chant and shake rattles. Frequently, the dancers add percussive effects to the accompaniment by special musicians. They stamp on the ground, on the floor, or on a resonant platform with bare feet, boots, or high-heeled shoes, sometimes in complex counter-rhythms. Hungarian men click spurs; Russians click the heels of their boots as they leap. Austrian and Bavarian *Schuhplattler* males swat various parts of the anatomy in set rhythms. Sword dancers click swords: stick dancers click sticks in Spain, Portugal, England, Mexico, Brazil, and India. Andalusians punctuate their incisive foot rhythms with crisp sounds of finger castanets; Greek males click spoons in their *zabakelos;* and American Indians sometimes shake rattles. In such secular dances as the Cuban rhumba or Argentinian *carnavalito,* accompanists use rattles. Sound makers may be attached to the costume, as the bell pads of Morris dancers or the ankle bells of India's nautch dancers. In many parts of the world exuberant dancers dispense with instruments and clap or shout at specified times or whenever the spirit moves them. They may also sing to various instruments or without instrumental accompaniment.

Percussive effects produced by dancers

Song. Self-accompaniment by song is significant for several reasons. First, it is probably one of the most ancient forms of accompaniment because of the independence from any instruments. Second, it is aesthetically pleasing. Finally, the songs have texts of historical, sociological, or ecological importance. Such singing may be in unison, with women's voices an octave higher than the men's; it may employ harmonies characteristic of the region, with intervals of a third or a fourth, and it may involve antiphony between a leader and the dance group or two groups of dancers. Such antiphony occurs in widely separated parts of the world, frequently in connection with serpentine chain dances as in Manipur and in America's Southeastern woodlands. Frequently the responses use nonsense syllables, and they may involve gestural responses, as in the Cherokee "stomp dance" and its predecessor, the ancient Aztec serpent dance.

The song texts are varied. The most frequent topics are courtship, as in the "Llorona" of Mexico's Tehuantepec, or sheer joy, as in the German "Freut euch des Lebens." In the Faeroe Islands the topics are narratives from legends, which are mimed by the round dancers. Sometimes the topic refers to agriculture, as in the French Canadian children's round, "Avoine" ("good grain").

The previous remarks mentioned the sound-producing items of costumes, as boots and bells. Visually effective items worn by dancers include kerchiefs and female full skirts, which permit numerous manipulations. Other visual effects are the designs of regional dress in the various countries, from the flouncy flamenco skirts to the white trousers of sword dancers. In the United States square dancers sometimes affect full skirts for women and plaid shirts for men.

CONTEMPORARY TRENDS

Revivals. The revived interest in national folk dances is generally dissociated from tradition, unless a folk dance group has a leader with folkloric knowledge. Folk dancing inspires the weekly gatherings of groups in civic centres, colleges, and other centres, even the entire schedules of summer folk-dance camps. Congresses sponsored by the Folk Dance Federation of California produced a uniform repertoire for groups throughout North America. In addition, new immigrants introduced occasional new dances. Most of these groups dance for the sheer pleasure of dance, and more expert ensembles stage programs and enter contests, both in the New and Old World. But although such revivals and the consequent preservation of traditions were heartening and brought about good fellowship, healthful exercise, and, avowedly, international understanding, such dancing had no connection with the aboriginal purposes of folk dancing, which continued only in villages or on Indian reservations.

Role of costuming

The modern style of costuming is an exteme departure from the masks for spirit impersonators and the symbolic designs painted or woven on all costumes of the ritual dances. Such paraphernalia survived in some dances that straddle ritualism and folk dance, as the animal and corn dances of the Pueblo Indians. But the trend was increasingly toward contemporary dress. Even the Iroquois ritualists usually wore ordinary clothes. Members of folk dance clubs rarely wore traditional costumes at their informal gatherings, although these clothes were customary for staged programs.

As a contrasting trend, professional folkloric troupes exaggerated costume effects, doubled the volume of skirts, added spangles, and increased the instrumental volume and the tempo. Frequently the directors composed scenarios, as in the reconstructions of Aztec rituals by the Ballet Folklórico de México. Their spectacles have a great audience appeal, compensating in part for the nonkinetic and the prosaic in modern folk dancing.

Continuity and change. The folk arts are by and large expressive of traditions that are deeply rooted in the life-styles and in the social organizations of peoples and cultures throughout the world. But as those styles and organizations change over time, in response to environmental, economic, technological, and other factors, so do the concomitant artistic expressions evolve in terms of function, form, and mode of existence. But change has been brought about, too, by the creativity of individuals and of cohesive groups.

Influence on other forms of the dance

Professional dancers found folk materials a rich source of inspiration that they used in several ways. Authentic dances were intensified for the stage by such companies as the Philippine Bayanihan troupe and the Ceylon National Dancers. The sophisticated dance-dramas of India's Uday Shankar, who performed widely in the West, often contained folk dances. His work with the Russian ballerina Anna Pavlova in *Radha and Krishna* showed, too, the rich potentialities for East–West collaboration. A folk atmosphere can be evoked without using folk materials, notable in *La Malinche* by José Limón. Finally, seemingly incompatible styles were fused: Mary Wigman was among the first to blend the rather stark idiom of modern dance with the ornate and exotic styles of the Orient.

Although the origins of many traditional dances are lost in a nebulous past, the observed emergence of new forms may give clues to the age-old processes of change. Inspired individuals may have molded the patterns of the ancient round-dance figures much as numerous leaders of dance in the 20th century have invented variations on the steps or devised steps and patterns to fit new rhythms, passing on their innovations by teaching or imitation. Again, creators have developed entire new dance structures from traditional materials, as the choreographers of modern Israeli dances have done most skillfully.

Another inevitable process is that of crystallization. For various reasons—sanctity, nostalgia, or whatever—groups tend to maintain routines through time. But not forever. If a dance does not die of old age, of having totally outworn its function and of having a form or spirit out of tune with a new age, it will continue to gain new life from improvised variations on basic steps or ground plans or from conscious elaborations of its forms by professional directors of ethnic dance groups and programs. Such kinds of creativity, individual and group, contribute to that constant cycle of orderly change within traditional parameters which accounts for the rich variety of the dances of the people. (G.P.K./Ed.)

Visual arts of the folk tradition

In the broadest sense, folk art refers to the art of the people, as distinguished from the elite or professional product that constitutes the mainstream of art in highly developed societies. The term in this comprehensive context combines some quite disparate categories of art; therefore, as a workable field of art-historical study, folk art is generally treated separately from certain other kinds of peoples' arts, notably the primitive (defined as the work of prehistorical and preliterate peoples). Historically, the terms folk and popular have been used interchangeably in the art field, the former being specific in English and German (*Volkskunst*), the latter in the Romance languages (*populaire, popolare*); the term folk, however, has increasingly been adopted in the various languages, both Western and Oriental, to designate the category under discussion here.

Difference between folk and popular art

Currently, the term popular art is widely used to denote items commercially or mass produced to meet popular taste, a process distinguished from the manner of the folk artist, who typically creates by hand (or with limited mechanical facilities) within a prescribed tradition objects designed for use by himself or his own circumscribed group. The distinction between folk and popular art is not absolute, however: some widely collected folk art, such as the chalkwares (painted plaster ornamental figures) common in America and the popular prints turned out for wide distribution, may be seen as the genesis of popular art; and the products and motifs long established in folk art have provided a natural source for the popular field.

Although the definition of folk art is not yet firm, it may be considered as the art created among groups that exist within the framework of a developed society but, for geographic or cultural reasons, are largely separated from the cosmopolitan artistic developments of their time and that produce distinctive styles and objects for local needs and tastes. The output of such art represents a unique complex of primitive impulses and traditional practices subjected both to sophisticated influences and to highly local developments; aside from aesthetic considerations, the study of folk art is particularly revealing in regard to the relationship between art and culture.

As industry, commerce, and transportation begin to offer all people free access to the latest ideas, and products, a true folk art tends to disappear; the integrity and tradition that formed its inherent character decline, and the heritage of home-produced products is undervalued for the very qualities that made it distinctive. Subsequent revivals, extensively sponsored by organizations, craft groups, governments, or commercial enterprises, are no longer the same thing.

The recognition of folk art as a special category came about during the late 19th century and was at first limited to the so-called peasant art of Europe, the "art of the land." The new intellectual climate of the time, with a romantic value attached to the simple life and the "folk soul" and the increasing spread of democratic or nationalistic ideas, brought the art of the common people into focus. It was recognized that their simple tools, utensils, and crafts had aesthetic aspects. Prior to industrialization, such folk art was widespread throughout Europe, exhibiting almost everywhere local styles created by people who had no access to the products of the wealthy and who were engaged largely in agricultural, pastoral, or maritime pursuits. As sophistication advanced, localism began to break down along major routes, but the folk arts continued on the periphery, particularly in geographically isolated regions, where they had an opportunity not only to survive but also to elaborate.

Preservation of folk art

Having only limited contact with the outside world, the inhabitants preserved their traditions, art forms, and methods of workmanship over a long period and, at the same time, had to rely on their own invention to create new styles and products at need. These outstanding regional arts provide a well-defined core of material in the field of folk art.

As the early colonists emigrated to remote parts of the world, they, too, were isolated from the cultural developments of the homeland and forced to rely on their own skills for most of their products. The arts they took with them were transformed, and new arts emerged under the stimulus of a different environment and through contact with native cultures; the notable folk arts of the Americas were one result.

In time, it was recognized that the great Oriental civilizations, like those of Europe, also had two distinct forms of art—the elitist and the folk. As Oriental folk-art scholarship developed, the subject gained international footing.

While most scholars agree that a folk type of art has occurred at some time in many parts of the world (and may yet appear in newly developing countries), there are various areas in which such art has so far been ignored or has not been studied as a separate category. For instance, with the notable exception of Roman folk art, the folk distinction is not usually applied to the art of ancient civilizations nor to Islāmic or Western medieval art. The summary provided here is, therefore, necessarily concentrated on the more studied areas: European folk art of the 17th–19th century, colonial and postcolonial folk arts, and the folk art of certain major Eastern countries. In addition to the major folk regions, this article will deal with the categories, styles, content, and motifs of folk art.

PATTERNS OF DEVELOPMENT

The extensive studies of European and American folk art over the past century have revealed certain patterns of folk-art development. Though these patterns are subject to revision as the field expands or is refined, they provide a basis on which cultural variations and less widespread or random occurrences may be considered.

The utilitarian aspect of folk art. Typically, the people who created the art were immediately concerned with producing the necessities of life; as a result, the art is often described as predominantly functional or utilitarian, in spite of the fact that there are important categories that are definitely not utilitarian, such as the widespread miniatures created simply for pleasure. It is true, however, that much artistic effort was absorbed in meeting everyday requirements. In the folk group, in which occupations were often seasonal or dependent on weather and where people had to provide their own amusements, the creation of useful objects became also a leisure-time activity on which creativity was lavished; a shuttle might be transformed with carving or a chest with painted designs, and even the corset stay came to be an art form. For this reason, folk art is best studied (as is primitive art) with the entire handmade product included and attention devoted to its cultural as well as its aesthetic significance. It differs from the study of sophisticated art, in which there is a long-standing distinction between fine and applied arts and a tendency to exclude, or at least segregate, the utilitarian from more strictly aesthetic forms.

Folk art was not created for museums. Certainly, some was designed to endure, such as documents, family portraits, and gravestones; occasional types were made purely for display, such as the "show towel" of the Pennsylvania Germans and the sampler (a piece of needlework with letters or verses embroidered on it as an example of skill); and certain household treasures were preserved for generations. In general, however, there was an indifference to permanence, so long as the function was served; and much of the art was expected to be either consumed or discarded after a celebrative appearance. There is a substantial percentage of intentionally ephemeral folk art—the marriage bowl broken after the ceremony, paper objects burned at funerals, festival breads, carnival figures, graffiti, snowmen; temporary symbolic designs were drawn on the threshold on feast days in India, for example, and were formed of flower petals for religious processions in Italy. Folk-art collections, thus dependent at least in part upon the accidents of survival, must be supplemented by photographic and written documentation in order for a representative view of the whole art to be obtained.

Art to be destroyed

The role of continuous tradition. The element of retention (prolonged survivals of tradition) is considered fundamental in folk art, as it is in folklore. In an isolated situation, the sophisticated ideas that penetrate are generally belated and simplified, and there is a natural trend toward conservatism. Both local and ancient traditions maintain a strong hold. Serviceable forms and familiar motifs are likely to persist, and changes are gradual in comparison to the sudden innovations possible in sophisticated art.

Yet a constant individuality and ingenuity affect the familiar mode, and an art uninhibited by arbitrary aesthetic rules takes many fresh directions. Thus, the fluctuating combination of retained and inventive elements is of significant interest.

CHARACTERISTIC MATERIALS AND TECHNIQUES

The most easily distinguished characteristics of folk art as a whole relate to materials and techniques. Most commonly used were the natural substances that came readily to hand; thus, various materials that have little or no place in sophisticated art, such as straw, may figure importantly in folk art. Sophisticated media, such as oil painting, might be adopted if they could be manipulated, and manufactured products—notably paper, which was cheap and versatile—might be used where available. The unique forms evolved in these sophisticated media illustrate the way in which folk art draws upon the general culture in a limited way, while developing along original lines of its own.

Tools were usually few and often multipurpose: delicate Polish cut-paper designs were often executed with clumsy sheep shears; and in woodwork, chip carving (with ax or hatchet) and notch carving (V-shaped cuts with a knife) were widely used.

Some arts were well within the compass of folk technology; textiles often rival the sophisticated handmade product in workmanship (differences being a matter of styles and themes). In many crafts, however, the folk artists evolved simpler methods of their own. Cut tin, in silhouette shapes or decorated by hand painting or pricking (marking out a design with small punctures), for example, is a common folk medium, whereas full-round bronze sculpture was not likely to be attempted. Again, the French Canadians used wood for "cathedrals" that were carpentered adaptations of their European stone prototypes.

Devices used to overcome technical deficiencies

Large-scale figures often reveal special devices that were invented to overcome technical deficiencies; some are crudely assembled from parts; many maintain a simple overall shape with details merely incised; feet might be represented by pegs inserted into bored holes. In pictorial representation, the difficulties of three-dimensional modelling, while readily solved by some groups, frequently resulted in a preference for outline and flat shapes; for the easier, profile view; and for the evolution of such forms as the silhouette and the shadow picture, made by outlining and filling in the shadow of a head cast onto the wall or paper. The limitations forced a mutation in forms.

Folk art in the urban environment. Folk art is by no means restricted to characteristic regional groups or rural arts. It occurs, for example, among minority groups bent on preserving their ethnic or religious traditions and their typical products. There are various folk manifestations within an urban environment, particularly in connection with the celebrative arts, which have a strong traditional hold; for example, at Christmastime in Warsaw, the people carry about the city models they have made of their cathedral. Covered with salvaged coloured foil, the models incorporate a Nativity scene and are lighted by candles or, more recently, by small bulbs and batteries.

Collective versus individual art. While many folk artists are known by name and many specialized in a particular art form, the skills were mainly available to all (with a distinction between the crafts of men and women), and most of the people were productive. The originality that delights the collector was not emphasized by the people themselves, who were concerned with producing the best examples they could of the desired object decorated with the appropriate and traditional image. Without consideration of the group involved and of the circumstances of folk culture in general, the art can scarcely be interpreted.

CATEGORIES OF FOLK ART

Only a part of folk art falls into the recognized sophisticated categories of visual art, and even that part has its own adaptations.

Architecture. In architecture the focus is naturally on the basic dwelling and on a simple public or religious building. One of the oldest and most remarkable dwelling forms survives in the *trullo* of Puglia, in Italy. A circular dry-stone structure with a tall conical roof, it is often decorated with symbolic designs splashed in white; for multiple rooms, the basic construction is simply repeated. The whitewashed stone architecture of the Greek islands, combining basic cubic forms with a variety of free shapes and inventive projections of balconies, overhangs, and exterior stairways, has been extensively studied and acclaimed by modern architects—as have the wooden churches of eastern Europe, with their delicate, needlelike wooden spires, and the wooden-stave churches of Scandinavia. Other unique forms are the Alpine house, with its steep, wide-eaved roof designed for snow; the cave dwellings of Spain, some with several rooms and a constructed exterior front; the adobe house; and the log cabin. A characteristic design may evolve for such outbuildings as the granary (notably the *hórreos* of Galicia), the dovecote, the straw shepherd's hut, or the barn. In community building, the walled agricultural villages with radial pathways to surrounding fields, the fishing villages which are oriented to a harbour, and the American stockade cluster as well as the village common exemplify the close relationship of folk design to folk activities.

Painting. The idea of a picture to be hung on the wall is by no means universal in folk art. It occurs in Eu-

Street on the Island of Mykonos, Greece, Greek island folk architecture.

Myron Goldfinger New York

The "picture" in folk art

rope, notably as the ex-voto, or votive offering, hung in churches and chapels, and in America, where portraits and local scenes were executed in oil, pastel, or watercolour. More typically, the painted depictions that occur in folk art are incorporated into other objects; for example, the American clock faces bearing local landscapes. A feature of some folk art is the "picture" displayed like a painted one but executed in such nonpaint media as fern, cork, shells, or embroidery. Oil paints and prepared canvasses are sophisticated materials and, though sometimes available, were often replaced by house paint or chalk and by silk, linen, or cotton fabric. Painting on velvet and underglass painting emerged as specific folk types. The amount of decorative painting on a particular object is often very extensive; among German and German-American groups, for example, every inch of a chest, bed, or chair surface might be covered. Walls or beams were commonly decorated with geometric and floral motifs and occasionally with scenes, though the available space did not encourage anything approximating the sophisticated mural. Painting on exterior walls was a feature in some areas, including parts of North Africa and India as well as Europe. Stencil painting, widely used for furniture and walls, illustrates the folk capacity for achieving varied effects within technical limitations.

In America the technique was applied to "theorem painting" (painting on velvet through a stencil, usually done with a dauber or pad and with some attempt at shading).

Miniatures

Sculpture. Some form of figural sculpture and a quantity of incised or relief decoration applied to a variety of objects appear to be almost universal among societies. Work in wood was particularly widespread, though stone, a more difficult material, was also used, especially for gravestones and religious sculpture. Papier-mâché, with its quick and bold effects, was widely adopted both in the East and West for carnival and votive figures and for a multitude of toys. The folk artist was often at his best in making small things, delighting in toys, small-scale representations of daily activities, and such oddities as ships carved inside bottles. Miniature sculptures were often skillfully executed in elaborate groups displaying a cohesive harmony; in Russia, for example, an entire herd of cattle was mounted on a jointed trellis designed to provide a scissorlike movement to the whole. Some figural types were created to be set up in groups, as were the European crèche figures (making up the Nativity or manger scene), toy soldiers, and Chinese miniature wedding processions. The creation of useful objects in an overall sculptured shape, both in pottery and wood, is also typical. In southern Europe or in Mexico, a bottle, flask, or candlestick might take human, fish, or other forms; a Moravian beehive, for example, might be a sculptured head.

The folk print. The wood block (also used for stamping textiles) was the natural folk medium for making prints. Usually simply cut and sometimes crudely coloured or stencilled, they served to illustrate popular subjects, with more interest often in the idea than in the depiction itself. Small prints of various saints were widely produced in Europe. Comic themes were popular, such as the "topsy-turvy world" and "man reversed" (*e.g.,* "the fish catches the man") and stock characters. Block printing was also used to produce games, announcements for travelling shows, and forms for certificates. The English broadsheets and the Mexican *calaveras* (literally "skulls," a category of prints, sometimes made from lead cuts) offer outstanding examples of the cheap printed sheets that combined a verbal message (verses, proverbs, polemics, pious themes) with illustration. The 19th-century trade cards (notice for a shop or service) are sometimes included in folk art, but doubtfully so; they were often machine printed. In fact, it is difficult to segregate the print of truly folk character from the voluminous field of either "popular" or commercial printing.

Other arts. In the folk field, the minor arts can hardly be called minor, for such universal necessities as pottery, textiles, costume, and furniture and more unusual forms such as weather vanes and scarecrows provided the most frequent opportunities for creative expression and often absorbed the aesthetic impetus that, in the sophisticated world, was associated more with the fine arts.

Both pottery and textiles range from the everyday to elaborately decorated forms that are often symbolic or highly pictorial; even common examples are typically ornamented with design in a simple slip (a mixture of clay and water) or a woven band.

Folk costume is justly included in many general works on costume, but it differs significantly from the sophisticated in several respects: in a localism so extreme that even a particular town or valley may have its own prized style and every region is distinctive; in the complete differen-

From the Scrimshaw collections at Mystic Seaport Mystic, Connecticut

Scrimshaw, whale tooth inscribed with sea chart by John Rause, 1790. In the White Collection, Mystic Seaport, Mystic, Connecticut. Height 15.5 cm.

tiation of the festival costume from ordinary clothing; and in a prolongation of style that is little affected either by changes of fashion or by individual taste. The motifs which are typical of festival costumes, such as the twin, cone-shaped buttons symbolizing fertility in Sardinia, are too deep-rooted in the tradition of the area to be discarded.

Furniture tends toward basic, repeated shapes, which may be left purely functional but are often extensively carved or painted. The Alsatian chair, for instance, has an upright-board back, carved with a pierced, silhouetted, bilateral design; some hundreds of variations of this simple design have been recorded within the area. Certain occupational forms emerged, according to need, such as the milking stool, the cobbler's bench, and the rocking bench, or "mammy settle."

In metalwork, the materials used to produce tools and other essentials were also turned by the craftsmen into such art forms as toleware (painted tin or tinned iron), incised copper or silver, pewter toys, and lead figurines. European wrought-iron grave crosses and shop signs are distinguished by intricate scrollwork and inventive linear depictions. Delicate bone carving is very widespread, appearing on such objects as implements, game pieces (such as chessmen), figures (notably crucifixes), and ornaments. An art peculiar to North America is the whalebone carving (scrimshaw) made by sailors while at sea.

Theatrical arts

The theatrical arts are spectacularly represented by puppetry, ranging from toy theatres, finger puppets, and the ubiquitous Punch and Judy shows to the famous puppet theatres of Sicily and Indonesia. Among the appurtenances of travelling shows and miracle plays, dating from the earlier phase of European folk art, was the hobbyhorse, which had a counterpart in festival performances in India. Musical instruments offer a profusion of types, often preserving ancient features of construction, principles of sound, and decoration: the heavy ratchets and rattles of the Alpine festivals; the shaggy bagpipes of the Abruzzi mountains; fiddles such as the rudimentary *gusle* of Yugoslavia, with its typical horsehead or horseman scroll, and the more complicated Norwegian *Hardangerfele,* with underlying sympathetic strings; and innumerable ornamented flutes, harps, horns, and dulcimers. The simple, painted clay whistle or flute is widespread, often in mimetic bird shape.

Specific folk categories. Any attempt to analyze folk art in terms of the established, sophisticated categories, though revealing in comparison, fails to take into account a substantial bulk of the art. Many characteristic products not subject to sophisticated aesthetic treatment have become specific fields of study and collection because of the ingenuity expended upon them—mangles (laundry beaters), molds, decorated eggs, weather vanes, decoys, powder horns, trade signs, scarecrows, and figureheads, to name a few. There are also significant objects categorized according to function; for example, animal gear represented by the woven harness of donkeys in Spain, carved and painted ox yokes and sheep collars, brass-studded and tasselled headpieces, and ornaments supposedly endowed with protective powers. Other widespread types are decorated vehicles such as gypsy and circus wagons, boats bearing symbolic motifs, and toys and miniatures in countless media.

Freedom of media. While some of the art is executed in a recognized sophisticated medium like wood carving, many other materials, such as hide, horn, straw, bamboo, and palm leaf, are characteristic in certain regions or for certain objects. In fact, there is scarcely an available material that is not utilized somewhere in folk art, from the hickory-nut doll to the commemorative picture made of human hair, and materials are often combined. This free-wheeling employment of any sort of material rivals the fertile adaptations of "found objects" in 20th-century sophisticated art—as many other modern "innovations" have a long-standing precedent in the spontaneous art of the folk. Collage, and assemblage are an old story in this field; embroidered pictures had faces painted in watercolour, and festival figures were made of anything that came to hand. Weather charms in southern Germany were often collages of—among other things—saints' pictures, amulets, and seeds. There is also a great deal of kinetic art: manipulated masks; jointed dolls, figures, and toys; whirligigs (spinning toys); pinwheels spun by wind or candle heat; and balance figures set in motion by a touch. Folk festivals, with their impromptu processions, costumed personages, antics, and props, offer almost a prototype of the 20th-century "happening."

Kinetic art

STYLE

Although the folk artist had his own criteria of function and craftsmanship, design in the theoretical sense was not a part of his training; rather, it was the natural result either of his continued use of established patterns or of instinctive methods of organization. In special instances there was deliberate imitation of well-known works of art, as in the American portraits of George Washington and folk versions of famous Virgins and Buddhas.

Any particular folk art will necessarily share the style of its general cultural area; Chinese folk art is Chinese as well as folk. Thus, analysis of the style and recognition of its folk origin is dependent upon knowledge of the "high art" with which it interacts, as well as of the folk situation that sets it apart. When a folk piece is compared with an adjacent sophisticated one produced at the same time, the differences become apparent, whether in the nature of the object as a whole or in its material, execution, content, or style. Stylistically, the time lag is significant; for example, the Baroque curve survived in simple country churches, and elaborate floral ornament in furniture decoration, long after sophisticated European art had become neoclassical.

Commonly accepted notions of folk style

One of the commonly accepted notions of folk style is that it is naive; it is thought to be childlike and fresh, despite the fact that some of its 19th-century critics condemned its "meaningless repetitions" and its "degenerate" forms. Repetitiveness is to be expected in the production of objects needed by all; but the artists saw only a few neighbouring examples, and to the practiced eye their art reveals many variations. Folk art is often associated with bright colour and an appealing charm, qualities sufficiently present to account for a wide popularity but counterbalanced by the sombreness and seriousness of many pieces, notably in religious art. In fact, few commonly accepted notions of folk style apply to the entire field. Execution may be free or meticulous. Representations of figures may be highly literal (even to the inclusion of actual hair and clothing), almost abstractly simplified, or monstrously exaggerated and distorted, as in, for example, the boldly painted papier-mâché carnival figures of Europe or the fantastic animal figures of the Far East.

The focus on utilitarian production leaves its mark in two opposite ways: often there is a strong decorative orientation, with a wealth of surface ornament lavished on objects that maintain a prescribed shape; on the other hand, certain categories of folk production, such as simple tools, and the work of certain groups are characterized by a functionalism so complete as to seem in tune with modern sophisticated design. Technical limitations and the demand for a quantity of certain necessary objects are conducive to simplification; the reverse may be true of such an object as the bridal bedspread, for which custom dictates extreme elaboration.

The particularly long retention of traditional forms and patterns generally results in increasingly stylized versions of themes; in crewel embroidery, for example, the representation of landscape elements is commonly reduced to a tree and hills, the hills typically shown as three simple, rounded humps; in American portrait painting, the bust or figure is conventionalized in a simple frontal form, repeated over and over again and sometimes painted in advance of a sitting, leaving only the features to be filled in. More important, perhaps, is the fact that the adoption of materials not used in sophisticated art, the forcing of a limited technology toward artistic expression, and the adaptation of rather remotely perceived sophisticated ideas to the folk artists' concept of the realities of life result in some highly original stylistic solutions.

CONTENT AND MOTIFS

Whereas sophisticated art often reaches out for the esoteric and the unusual, the content of folk art is closely related to immediate human concerns. The major events of life were universally celebrated on the folk level in ways that demanded of art special costumes, implements, vessels, and auspicious gifts. For the newborn there might be amulets and decorated birth certificates. The period of courtship occasioned a love token, often a beautifully carved feminine implement such as a shuttle or needle case; traditional in England was a double spoon symbolizing union and plenty, whereas in Czechoslovakia it was often a painted egg or carved stick. In many regions elaborate wedding chests were carved or painted for the bride. The bridal bedspread or bed curtain, like the wedding costume, was ornate and highly symbolic, with such motifs as Adam and Eve, the tree of life, and mating birds considered appropriate. Both weddings and funerals required processional equipment, standards, and special vehicles. In some places there were gifts for the dead, which in China took the form of paper models burned at funerals. There were memorials such as grave sculpture, pictures, and documents.

Specific memorial motifs crystallized in two American forms: the "mourning picture," executed in embroidery or watercolour, often depicting grieving figures draped around a tombstone under weeping willows, and the gravestone carved with a winged death's-head or, later, with the urn-and-willow motif.

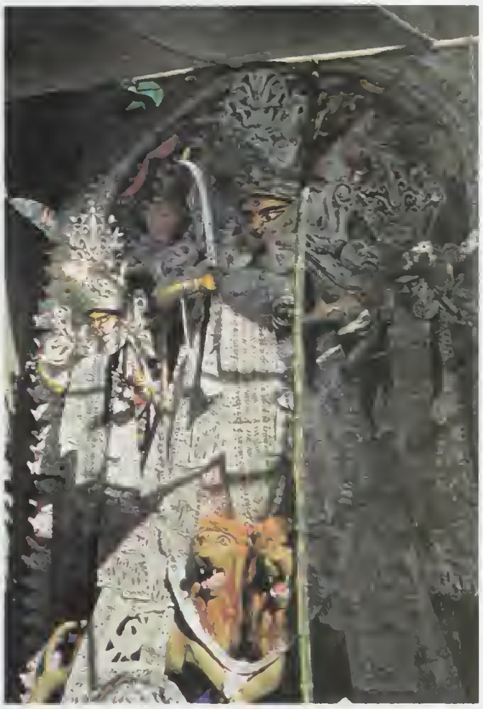
Religious art. The prevailing religion puts its stamp on the consciousness of every group, providing common elements in areas that share the same religion, even though the groups are not in contact. Catholicism in the West (and, similarly, Buddhism in the East) provided rich visual conceptions and evocative images that spilled over into folk art. Crucifixes, Virgins, and saints were required as images for village churches and wayside shrines; they were set up over gateways and tombs, in arches, and in homes and were used as motifs on countless objects, where they were often freely combined with secular decoration. Religious observance demanded many objects decorated with Christian symbols—baptismal scoops, altar cloths, pilgrimage bottles, lavabos (holy-water vessels). There is even a special category of "nuns' work," including small devotional objects, many in collage, as well as vestments and church textiles. A particular German sculptural type is the *Palmesel*, a half-size figure of Christ on the donkey, which is drawn through the streets on its wheeled base on Palm Sunday.

Influence of Catholicism

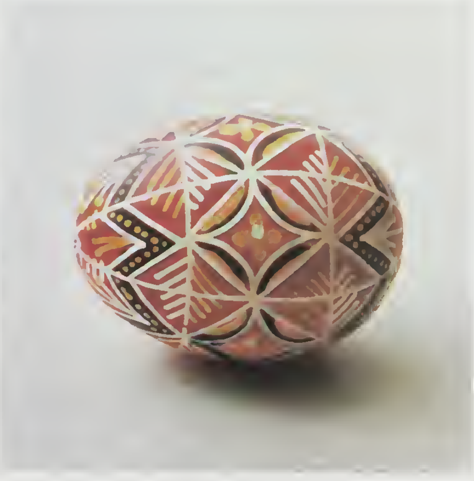
An outstanding category of Catholic folk art is the crèche, made up of figurines displayed at Christmas in homes or churches to reenact the birth of Christ. The main characters of the event (holy family, Magi, shepherds, and angels) were supplemented by hundreds of lively figures drawn from peasant or village life and shown pursuing their daily activities or bearing gifts to the Christ child similar to those enumerated in folk carols.

The Protestant and Jewish faiths made fewer demands on the visual arts, but the popularity of Biblical themes is apparent. A favorite motif for the American weather vane was the angel Gabriel blowing his trumpet, often executed in a style that survives from the puffing zephyrs of classical art. The noteworthy Jewish folk art of Poland was largely lost during World War II, though records of the unique folk synagogues have been preserved by the Institute of Polish Architecture. The Jewish folk-art collection in the Musée Alsacien, Strasbourg, France, includes such specific religious objects as pointers (carved sticks used to guide the reading of sacred texts) and candelabra.

Since antiquity, some form of votive art has occurred in connection with religion. In India, outdoor shrines may be surrounded by a veritable crowd of papier-mâché figures set on the ground as offerings. Catholic churches and chapels throughout the world are hung with countless small ex-votos, usually cutouts of stamped tin or silver in the shape of an afflicted part of the body—an arm, a leg, or an eye—or of the heart or other symbol. In Canadian Jesuit missions, ex-votos were even made of wampum. In Seville small ivory carvings of religious figures were left in the cathedral by soldiers going to war. Clay plaques made from molds, common in the Mediterranean area, show an inheritance from Greek times, when small clay molds of the head of Athena were stamped out in quantity as votive objects. The most significant art, however, occurs in the painted ex-voto, which provides a major type and some of the best examples of folk painting. In sophisticated art, paintings of standard religious themes were often donated to churches in fulfillment of a vow. In folk art, this votive urge found expression in small narrative paintings (only occasionally large, as in Mexico) depicting an accident, illness, or other disaster from which the victim was saved by the intervention of a saint or the Madonna.


Festival votive figures of Durgā, papier-mâché and tinsel, 1961, West Bengal, India. Height 5.2 m.
R F Bussabarger

The recognized religion, however, is only a part of folk belief, which is impregnated with concepts from earlier times. The decorated Easter egg, for example, is an evolution of the egg as an ancient symbol of renewed life, and the fat, laughing figure of the Japanese Hotei (god of luck) is both a deity and a ubiquitous folk charm. There are many survivals from local pagan cults, particularly of motifs associated with life, fertility, and protection; in Calabria an animal stake may be carved with the head of the blank-eyed mother-goddess, expected to protect the tethered beast, and similar elemental forms were preserved in Czechoslovakia. Lying at the root of human experience, such themes were never completely abandoned by the folk and may appear in curious juxtaposition with Christian


Decorated Easter egg from Czechoslovakia, 20th century.

themes or secular uses: a Sardinian clay bowl, for example, contains a modelled wedding group with the priest standing before an altar on which a small, nude hermaphroditic deity is seated, and the Christian loaves of bread appear along with pagan phallic and fertility symbols.

Ancient pagan seasonal celebrations

Festival art. A major folk category is festival art, which owes its genesis and much of its content to ancient seasonal celebrations. Since antiquity, the solar manifestations of the summer and winter solstices and the vernal and autumnal equinoxes have been bound up with the idea of sowing and reaping, death and rebirth, year's end and year's opening; at such times it was traditionally believed that supernatural forces were in control and should be propitiated. Re-enactment of the roles of malign spirits called for the production of grotesque masks and demonic costumes and also of clamorous noisemakers (bells, horns, rattles, and the like) to drive them away. Harvest figures invoked or celebrated a good crop yield. Special foods in symbolic shapes were prepared and consumed. Varying according to the culture, many other appurtenances were created—decorated trees and poles, lanterns, banners, processional vehicles, sculptured figures and dolls, household and shrine adornments—all bearing their motifs of life symbolism.

While the magical significance of the primordial festivals may have been largely forgotten and the events reduced to horseplay and merrymaking, the customs and the art objects associated with them persisted. In Europe, masqueraders continued to impersonate such "characters" as death, the devil, the goat, the old man, and the mischief-maker; their masks were often makeshift and ephemeral, but many carved of wood and decorated with other materials are preserved and highly prized. Such personifications were also painted on banners or created by assemblage and carried about, as were the Mexican skeletal death figures.

Oriental festivals often featured plant and animal motifs. In China the dragon of the New Year was a great paper creation made to undulate by the dancing steps of the bearers underneath. In the Japanese boys' festival, painted paper carp were flown from poles as symbols of strength and virility. In Indonesia, towering decorative constructions of vegetables and fruits were borne about to celebrate the harvest.

The assimilation of ancient seasonal celebrations—the winter solstice and the Roman Saturnalia with Christmas, for example—has been extensively studied in European folklore. In folk art, it occasioned an intermingling of pagan and Christian elements, enriched by many inventions created in an exuberant festival atmosphere and readily incorporating local and current themes. The celebrative instinct found expression also in many purely local festivals commemorating a local saint, historical event, or an episode in folk life, such as the setting out of the fishing boats or the onset of rains. In Japan alone, there were hundreds of such festivals.

Other sources of folk motifs. The traditional survivals that play so significant a part in folk art stem from other sources as well. Certain motifs diffused from the earliest cultures provided a repertory of stylized symbols to meet decorative demands; for example, the rosette (a disk divided variously into petallike segments), the rayed disk and the swastika (both associated with sun symbolism), the tree of life, the chimera and other fantastic beasts, and such human-animal combinations as the siren or mermaid. The extent to which such motifs retain their meaning or may become simply an appropriate decoration for a certain type of object (as the mermaid is for boats) is problematical, but there is undoubtedly a high symbolic content in the art.

Use of medieval and Renaissance themes

Some aspects of classical mythology fed into folk art, partly by way of later European sophisticated art, and many medieval themes remained popular; the Saracen of the Crusades is a figure that still appears as a Sicilian puppet and as a revolving target in tilting games. Early Renaissance conceptions of paradise and landscapes with stylized trees and towered towns oddly recur in 19th-century folk painting, sometimes imparting an esoteric flavour to a local scene. In fact, the body of tradition retained in folk art may be seen as growing or shifting from one century or one place to another. A folk version of the horse-and-rider motif, in typical profile view, served with only a slight change of uniform for both the Napoleonic and the American Revolutionary soldier.

Although themes may fall into disuse, they do not become obsolete so readily as in sophisticated art. Yet, folk art is not merely a repository for tradition; new themes constantly evolve from old ones or out of new circumstances. In the wine-producing area around Alsace, Bacchus astride a barrel became the common motif for carved bungs (the stopper of a cask), thus utilizing the classical Bacchus for a specific local commodity. In America, the Indian was widely adopted for weather vanes, trade figures, and other objects. Similar use was made of the personification of Liberty and the emblematic eagle.

By courtesy of the Shelburne Museum, Inc., Vermont

Sheet-iron weather vane, unknown artist, *c.* 1810. In the Shelburne Museum, Inc., Vermont. Height 1.3 m.

Geometric and plant and animal motifs

All decorative design draws heavily on geometric and plant and animal motifs. In the folk use of this material there is often such concentration on one or two motifs that they become strongly identified with the regional style, as the tulip is in Pennsylvania German art; there is also a tendency to attach a particular motif to a particular object, for which it is used repeatedly. The prevalence of animal themes reflects the importance of animals in folk life.

Aside from their frequent appearance as realistic depictions, miniatures, and design elements, some animals also have strong symbolic aspects: the snake, the horse, and the cock, for example, occur with varying significance in many parts of the world.

Representational and narrative art other than the religious is often devoted to local subjects: the family portrait, the individual farm or church, or a typical activity. In Switzerland a favourite theme was the *Alpengang*, depicting the transferral of the cattle to high pastures in the spring. Folk artists also drew upon legend, popular romance, history, and the more famous literary and visual art themes that reached them from the sophisticated world. In Sicily the deeds of Roland (Orlando), derived from the poetic accounts by Torquato Tasso and Ludovico Ariosto, were repeatedly painted and enacted in puppet shows. From history, the patriot Giuseppe Garibaldi was as popular in Italian folk art as George Washington was in American; and the Prince of Wales was a favourite figure for pub signs in England.

Account must also be taken of the folk capacity for satire.

The anticlerical humour of Italy has a folk manifestation in caricatures of impious monks and nuns. The Russians evolved stock figurines of the snobbish officer, the vain woman, the greedy merchant, the pretty girl riding on a rooster. The early prints of London and Paris had their lampoons, and Mexico had its effigies of personages who did not meet popular approval. Out of the slow exolutions typical of a strongly traditional art, there emerges an astute view of the human situation.

MAJOR FOLK REGIONS

The major recognized folk regions in most cases have been prolific in such crafts as textiles, pottery, and carving and in the production of implements and utensils; they also often have localized costumes. This common art output forms a broad basis underlying the more distinctive arts peculiar to particular areas. The material is so voluminous that most attempts at general survey are admittedly samplings.

General summaries are commonly organized by nation, a convenient expedient, because major collections are centred in great national museums and because folk art is often studied and promoted as part of the national heritage. However, the national summary divides some groups that are homogeneous, such as the Basque region in Spain and France; and it combines, under Italy, for example, such diverse arts as the Alpine and Sicilian.

Any effort to group regions for comparative study will most logically be based on such factors as the traditional retained sources, the prevailing religion, the nature of the related sophisticated culture, and the environmental conditions that affect materials and activities.

Western. *Mediterranean.* Viewed in terms of these four factors, the European folk arts of the Mediterranean area obviously have much in common. First, there was a direct transmission from ancient Near Eastern and Greek civilization, accentuated by Greek colonization in the West and followed by Roman domination. These sources, plus the local cults that occurred everywhere, may be traced even in recent art in the continuance of a rich pottery tradition from Greek times onward and in the preservation of many motifs. Second, the religion, chiefly Roman Catholic or Greek Orthodox, demanded extensive imagery. Third, in the sophisticated cultures throughout the historical period, art of all kinds was a major activity, developing high skills that penetrated to some extent even to the more isolated folk. Finally, contact was facilitated by active trade along an extensive coastline, and varied materials were available; yet the area industrialized very slowly, so that the folk arts could continue to thrive in some localities even to the present.

Outstanding quantity and variety of Mediterranean arts

Thus, it is not surprising that the arts of this region are outstanding in quantity and variety. The level of skill is apparent, sometimes in bold and facile styles, sometimes in meticulous craftsmanship. Many folk artists were capable of expert full-round sculpture, realistic painting, fine metalwork, and other difficult techniques. The motifs are varied and freely intermingled.

Among the long-surviving regional arts are those of Epirus in Greece, where an important folk centre has been established at Metsovon; the islands—the Aegean with its stone architecture, Sicily with its spectacular carved and painted carts, puppets, and pottery, and Sardinia, noted for gold ornaments, textiles, and costumes still in use; and Puglia, Calabria, and the Abruzzi region in Italy, the latter having fine lace, silver filigree (openwork), and weaving.

Southern France is affiliated with this area, as is evidenced by the style of the fine ex-votos and Nativity figures of Provence. So is the Iberian Peninsula, though in that region there are also special factors. The Moorish influence was felt particularly in Andalusia—as in the use of ivory as a material and in the arabesque tracery (ornate, interlaced openwork) of the ironwork—and the Atlantic coastline provided other connections. The Portuguese use of cork was distinctive. Galicia and the Basque region, each with a population of distinct linguistic background, developed in prolonged isolation, the results of which are clearly visible in their exceptional arts: the architecture presents unique features, and the Basques are unusual in their lack of pottery, though they developed remarkable dance costumes. The difficulty of communications preserved a strong folk character not only in Galicia and the Basque region but throughout the peninsula. The painted and glazed tiles (*azulejos*), the textiles (notable in Salamanca), and carved furniture are among the products notably Iberian in character. Traditional survivals were strong in the northeast, with much religious art, including prints, centred in Catalonia.

Slavic area. Another possible grouping is the Slavic area in eastern Europe and Russia. There the influences from the ancient Near East and Greece penetrated less far in early times but were transmitted (and transmuted) by way of the Byzantine Empire and the Eastern Church. Much folk art in the area was strongly affected by the Byzantine style.

Among the transmitted elements were the themes and styles associated with icons, which were commonly hung—at wells, for example—until the mid-18th century, when their production was discouraged in Russia and thus dwindled. Two centuries of Mongol rule introduced other traditions stemming from the East and marked by the so-called animal style. Finally, in modern times, these countries have mostly had Communist governments, whose policy includes promotion of the folk arts, organization of artists into cooperatives, and even the introduction of crafts from one area into another. Although this has been a stimulus to the study of folk art, it tends to blur the distinction between the strictly folk and the revived or commercialized product. Even earlier, Russian folk art was subject to extraneous influences in a way not typical elsewhere: in the 17th century, craftsmen were requisitioned from many parts of Russia to supply products for the national economy or to work on palaces, and they were also assembled around monasteries for prescribed output.

The Russian products probably best known elsewhere are toys—intricate constructions of wood or vivid earthenware miniatures. Some of the Vyatka toys are thought to be survivals of idols made for homes, representing the innumerable local deities that preceded Christianity. Other notable arts include ceramic tiles, wooden and ceramic figurines, and bone carving in the Siberian tradition.

By courtesy of the Walters Art Gallery, Baltimore

Carved walrus-bone comb from northern Russia, 17th century
In the Walters Art Gallery, Baltimore. 7 × 13 cm.

Eastern European woodwork

In eastern Europe, where national boundaries have been particularly confused and the population comprises various minorities, studies of the art may follow ethnic lines. The geography provides a number of distinct regions, which are as varied as coastal Dalmatia, Transdanubia, and the isolated Tatras mountains. With a heavily forested landscape, the work in wood was outstanding. It appeared in church architecture, architectural sculpture, vessels and implements, and in such special forms as the sculptured grave-post; even a corn bin might be covered with rosettes. The area was rich in festival arts, with a strong retention from pre-Christian traditions and magic rites. In Czechoslovakia there were special wedding effigies and candlesticks. Among many ancient motifs, such as vase-and-tree, sun, and heart, the cock appeared as a protective symbol that might be set on roof ridges or carved on cheese molds. Some of the art is strikingly primitive.

One of the complications arising in the study of eastern European arts is the fact that the countries involved are culturally borderline, having an affinity with Catholic

Europe in the West (exemplified by the ex-votos in the brilliant Czechoslovakian glass painting) and with the Byzantine Empire in the East. The arts bracketed as Polish, including some of the finest decorative art in paper, once extended far to the east and yet are northern European.

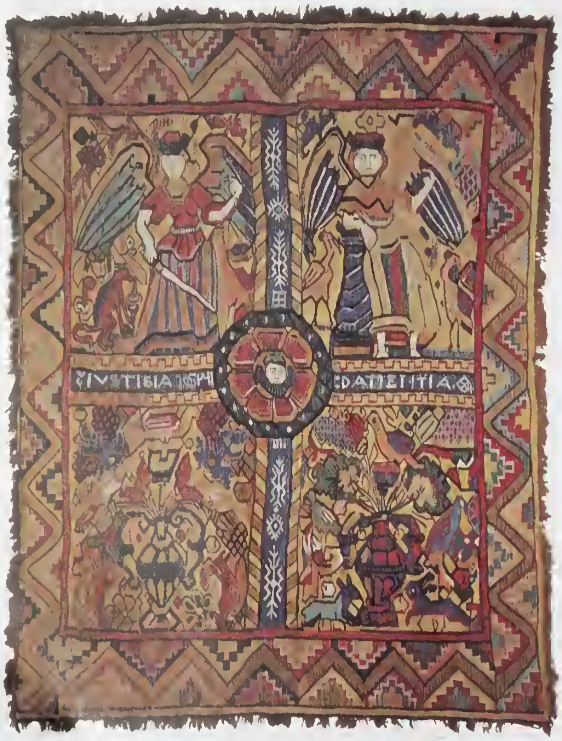
Northern Europe. The situation in northern Europe was very different from that in the south, and not merely in climate. The tradition involved a different mythology, and the society lacked the sophisticated centres that had crystallized early in Greece and Italy. The Roman influences that reached northern Europe had far to travel; consequently, the transmitted motifs were fewer, and emphasis might be placed on technical execution rather than on variety. This can be seen in the prevalent and superb use of two motifs, the acanthus and the vine-and-tendril. It can also be seen in the animal style from the East, which penetrated and persisted, for example, in some fine architectural carving, with the tendency typical of this style toward flat and pierced rather than full-round rendering. Finally, although religious art was by no means lacking, the Reformation, which in itself was a popular movement, curbed the use of extensive Catholic imagery as well as the demand for religious objects.

The festival arts drew heavily on northern pagan themes, and the impulse that gave rise to pre-Lenten carnivals of the south was likely to find expression rather in municipal and occupational processions with comic giant figures drawn through the streets.

Some parts of the far north demonstrate that density of population is a factor in folk art; where farms are many miles apart with few opportunities for community contact, the art forms may tend to be few or even nonexistent. Even so, there may be one or two special crafts, such as the bone and horn carving of Lapland. Also, where materials are scarce, as in Iceland, variety of product depends on imports likely to be allocated to the sophisticated, not the folk. In more densely populated France, Germany, and The Netherlands, on the other hand, it is clear that peasant arts existed everywhere in the earlier periods but that the early establishment of trade routes and urban centres pushed the folk arts into special categories or into the peripheral areas.

Among the Scandinavian regions, Norway is noted for the rose painting of Hallingdal and Telemark Fylke, the needlework of Hardanger, and the pictorial weaving of Gudbrandsdalen. Sweden, among varied arts, had a unique type of built-in furniture and wall hangings that were either painted or woven with biblical and Icelandic motifs. Finland had a specific linear ornament called "dark drawing," made by bending a strip of wood until the ends meet, and metal ornaments of prehistoric origin in Karelia. Distinctive folk-art regions in Denmark include the Hedebo (now Hedeboryde) area, with its linen embroidery; the Fyn archipelago, with its colourful floral painting; and Jutland and Slesvig, with notable cabinetmaking. In the Baltic area there were many survivals of ancient motifs (swastikas, rayed disks, snakes, horse heads) used on varied products, including the remarkable crosses and roofed poles, often with symbolic wrought-iron finials (crowning ornaments).

By courtesy of the Norsk Folkemuseum, Oslo



"Justitia," Gudbrandsdal tapestry, probably late 17th century. In the Norsk Folkemuseum, Oslo. 145 × 178 cm.

Central Europe. In the heart of Europe, two areas demonstrate special factors involved in the formation of folk culture: the Rhineland, where wine production provided a number of special objects and motifs; and the Alpine regions, which, though extending into several countries, share a pattern of living dictated by the mountain territory. The latter region, which includes several well-defined areas—such as the Appenzell in Switzerland, the Tirol in Austria, and the Alto Adige in the south Tirol, now a part of Italy—is rich in festival arts, ceremonial foods, and implements associated with dairying (even musical cowbells).

Arts of the Alpine region

In France, The Netherlands, and Germany, the proximity of folk groups to sophisticated culture made its mark in the variety of products, high skills, and lavish decoration of such objects as furniture. Invention was devoted to new figural types, such as the hod carrier common to lower Germany and Austria; and events such as the Napoleonic Wars made a rather quick impact, as with the soldier motif and the appearance of handwritten and ornamented documents relating to military service. The mechanical genius that made the Germanic peoples leaders in the field of sophisticated automata found folk expression in innumerable animated toys, clocks, chimes, figures, and other gadgets. While the folk art associated with Paris itself is not to be ignored, the more easily analyzed French groups are outlying, as in Brittany, with its many-figured outdoor calvaries (representations of the crucifixion) and other enduring forms.

Britain and Ireland. The tendency to separate British from other European folk arts is an oversimplification, for a number of forms are shared with northern Europe; for example, the famous horse brasses (circular harness ornaments often retaining ancient protective motifs), giant processional figures such as the Salisbury dragon, and the May tree, a celebrative decoration in pole form. England is a small country that industrialized rapidly, a factor that tends to shorten the folk-art period. Some arts that required expanding technical skills, however, could develop as folk forms: for example, the printed arts (such as the broadside, or sheet printed on one or on both sides and folded) and the hand-propelled roundabout (later the mechanized carousel), which became increasingly elaborate. Tunbridge woodwork, of glued coloured strips, is merely one example of local invention. Among the well-known categories of folk art are the inn signs (both hanging and "effigy" signs), wroughtiron work, and tombstones. Hebridean textiles and Highland plaids and sporrans (the pouch worn in front of the kilt) are also familiar products. Both Scotland and Ireland have interesting grave crosses bearing ancient symbols. Ireland, however, serves as a reminder that the creative urge of a folk group may not focus primarily on the visual arts; Irish folk art does not compare with the contribution to oral lore in that area. (The same may be said of the black folk minority in the United States, whose musical contribution was spectacular but whose visual art traditions were largely cut off.)

North America. The colonization of the Americas in the 17th and 18th centuries, stemming largely from Europe at a time when European folk art was flourishing, resulted in a second general area of major folk-art development. This art can be divided into that of the United States (loosely called American folk art); Canadian folk art, which has much in common with that of the United States, with its scrimshaw, ship carving, and western pioneer art, but which also has products of its own (for example, French-Canadian wood carving in Quebec); and Latin-American folk art, which has quite a different character.

For the first century and a half, the art of the eastern seaboard of the United States may be described as predominantly folk. Although there were European imports and works produced by sophisticated American artists, they were generally a pallid reflection of the art then developing in Europe, and they made little impact on a people bent on making a home in a new world. The so-called Yankee ingenuity produced a wealth of material, sometimes reminiscent of European prototypes but often new. There were, for example, dozens of handmade lighting devices and many specialized contrivances such as the trammel, for raising and lowering pots in the fireplace, and the corn planter. Fresh decorative styles, special forms, and new motifs contributed to an art that, either in evolution or invention, was typically American.

American folk painting

American folk painting is outstanding. Although there was once a tendency to view as sophisticated the artists who closely followed European styles and as folk those who worked in the rapidly emerging American manner, many of the latter have become individually known creators of a valuable body of work and have taken their place in the history of American art, some no longer viewed as belonging in the folk category. The more typical folk product comprised thousands of portraits and scenes by anonymous or local craftsman-artists or itinerant painters, who provided a vivid, if often crude, extensive record of America's ancestors and their surroundings.

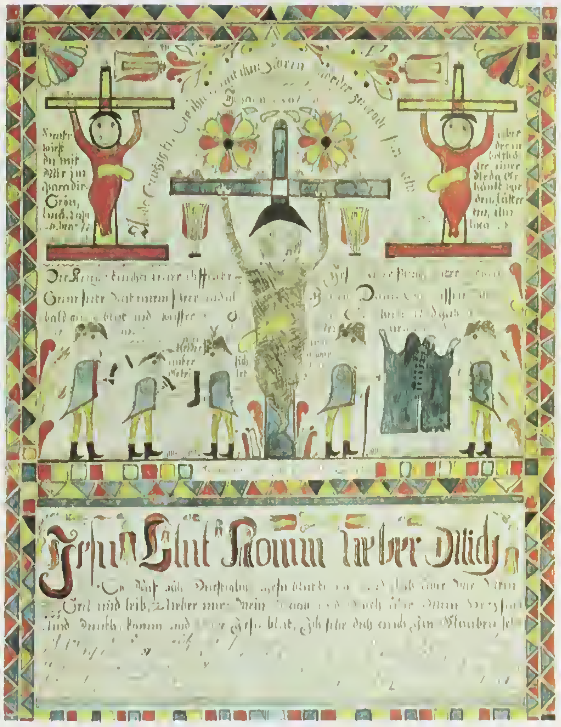
As America advanced, a pattern of regional differentiation appeared, just as in Europe. In general, geographical isolation was overcome rather quickly, one exception being the sparse settlements of the Appalachian Mountains, where Scots-Irish descendants maintained a handicraft tradition. People of varying origins, however, had brought to the "melting pot" of America their different art traditions. While they were often content with preserving a few objects and customs, some groups chose to maintain a separate identity, set apart by religion or national origin, and among them some fresh regional arts developed. The strict religious beliefs of the Shakers in New England and New York state, with their emphasis on simplicity, gave rise to clean, functional lines in furniture and architecture and to some psychologically interesting "spirit drawings" executed under the influence of religious visions. The Pennsylvania Germans (popularly called Dutch) not only had a distinctive religion but clung tenaciously to the language and traditions of their native Pfalz (Palatinate, now in the state of Rhineland-Pfalz and in Bavaria), which in art included such crafts as fine painted furniture and such motifs as the tulip, heart, and vine. Thriving in the flourishing countryside of their new home, they produced a notable body of art: fraktur (embellished documents), painted wedding chests, decorated ceramics (including elaborate pieces created for special occasions), unique barns with exterior painted symbols ("hex signs"), pictorial embroidery, weaving, and other forms.

Arts of the American west

The American settlers who moved westward were again thrust into a folk situation comparable to that of their forebears, and a pioneer art developed. Saddlery was one of its important crafts; the covered wagon was its distinctive vehicle; and the board structures of mining towns and the sod houses of the plains were solutions to the problem of immediate housing. The flatboat and keelboat of the Mississippi River arose from specific navigational requirements.

The southwest, including part of California, is an area apart, producing art distinct from what is often called "western Americana." There the architecture was influenced by the Spanish mission and adobe styles, and a Catholic religious art was encouraged among the natives, resulting in the carved or painted imagery of saints (*bultos* and *santos*) with a strong native flavour overlying the Spanish derivation. These arts are more allied with the Latin American (as may be seen in the Museum of International Folk Art in Santa Fe).

Latin America. The different character of Latin-American folk art may be ascribed in part to the modification of a primitive culture resulting from contact with an advanced one. The settlers on the eastern seaboard of North America moved in on a primitive Indian population whose arts were relatively limited and who were rapidly pushed back or disoriented; the folk art of that area was thus essentially the product of the white settlers. In Latin America, however, where there were some highly developed pre-Columbian cultures as well as tribal arts, intermingling was freer; this was partly because the missionary program—which included the teaching of crafts and Catholic symbols to the native population and the use of native craftsmen for church construction and for the production of religious objects—accepted an infusion of native techniques and styles. Thus, Indian crafts and motifs had a better chance of survival, and a greater degree of syncretism could occur. Furthermore, the colonizers, predominantly Spanish and Portuguese, brought with them the wealth of Mediterranean tradition and the varied imagery and forms of their home regions.


"The Crucifixion," fraktur by unknown artist, watercolour on paper, 1847. In the Abby Aldrich Rockefeller Folk Art Collection, Williamsburg, Virginia. 28 × 37 cm.

By courtesy of the Abby Aldrich Rockefeller Folk Art Collection, Williamsburg, Virginia

Explosion of folk art in Mexico

Under circumstances as favourable as these, a virtual explosion of folk art can occur, as it did, notably, in Mexico. Because Mexico seems to have a peculiar receptivity to art impulses regardless of source, the area is distinguished by a folk imagination that can create a towering, multifigured, ceramic candlestick, elaborate figures and models of straw, and fantastic fireworks. Craft motifs are handled with great spontaneity, and the festival arts are remarkable, with such original creations as the Judas figures, the skeletal musicians associated with the Day of the Dead (*dia de los difuntos*), and the skulls (*calaveras*) that appear both as confections and as a theme in popular prints. The religious arts are also outstanding, with many ex-voto paintings (*retablos*) and Nativity figures in varied materials. Art that combines features of the Mediterranean and native Indian traditions occurs also in other Latin-American countries, as in the Portuguese-oriented areas of Brazil and in Argentina, which developed some arts related to the life of the cowboy of the pampas (gaucho). In some regions of Latin America, however, the indigenous Indian culture long remained unaffected and little influenced by the European colonies.

In the Caribbean and coastal areas there is evidence of African–Indian–European interaction: saints are painted with African physiognomy, and African decorative motifs

Day of the Dead toys from Oaxaca, Mexico, pottery and paper, *c.* 1960. In the collection of the Girard Foundation, Santa Fe, New Mexico. Height of largest figure 26.5 cm.

By courtesy of the Girard Foundation, Santa Fe, New Mexico

appear on crosses, votive sculptures (the *milagre* of northeastern Brazil, for example), and such objects as laundry beaters and peanut pounders in Surinam.

Other regions. During the 16th–19th centuries, European exploration, trade, and culture expanded into many parts of the world. Colonization elsewhere, however, was not so conducive to folk evolution as in the Americas, where many settlers emigrated early, bearing folk traditions with them and expecting to make a life with their own skills. Because in many places the Europeans maintained a sophisticated enclave closely tied to the homeland, the native arts were preserved intact, inhibited, or exploited. This was fairly typical in Africa and the primitive Pacific areas, where settled colonization took hold only in the late 19th century. In South Africa, where it occurred earlier, only the Dutch (who built farmhouses of Dutch character) tended even to take their families with them.

Non-Western. In many parts of the world there have been tribal arts, some of which have nonprimitive aspects. These are sometimes bracketed with the primitive in a general category of ethnic art and are sometimes considered as folk art. But although they may have folklike crafts and links with the outside world, they differ from true folk cultures in that they constitute homogeneous societies with traditions that are specifically ethnic rather than shared with a broad area of sophisticated culture. Such tribal folk art occurs in the Saharan Berber and Siberian areas, among the Ainu people of Japan, and in various parts of Asia.

The Eastern art recognized as truly folk has been studied, as in the West, chiefly in the areas where it exists as the local or provincial art within a great culture. The Oriental traditions were often relatively uninterrupted, and effects from industrialization were late; while all folk dating is problematical and much of the art has perished, it is likely that some folk art in the East has a history extending back to ancient times. In Japan, however, it is usually understood as beginning in the Edo period (17th century). Interest in folk art is particularly strong in India and Japan, where there are art scholars familiar with the Western folk concept but dedicated to the preservation of their Eastern traditions. Indian folk art was discovered in an emotional climate reminiscent of the European discovery of the folk soul; Ananda Coomaraswamy, a leader in the movement, called folk art the "main road," as distinguished from the sophisticated "bypaths." Both in India and Japan, there were sophisticated artists who deliberately identified themselves as "folk."

India and Pakistan. In India, where all of the crafts are distinguished by variety, skill, and a strong component of strictly Indian tradition, the folk distinction is not always clear-cut. It is most apparent in such objects as toys (for example, the mother-and-child figure probably related to fertility concepts), masks, works in papier-mâché (votive and animal figures, for example, and dancing dolls balanced on wire), the symbolic motifs painted on the houses of the poor, and other works of art related to local custom or primitive belief. Particularly in southern India, small religious and other sculptures were created in quantity in an unmistakably folk manner; there are also some distinctive tribal arts, notably those of Assam. Pakistan has some highly regional arts: for example, the fine house carving and the ceremonial fans of Swāt, the silver ornaments of Gilgit, and the tombstones and matrimonial objects produced in the arid regions of Baluchistan.

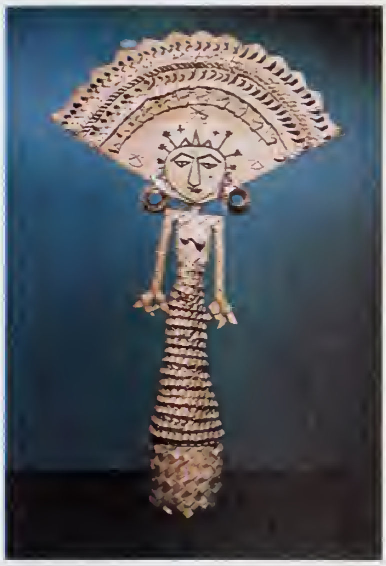
Pakistani regional arts

Japan. Pottery and toys are probably the most widespread kinds of Japanese folk art; but there are also innumerable typical objects—lanterns, fans, umbrellas, nested boxes, and kites—exhibiting skillful use of bamboo and paper, as well as wood, lacquer, and other materials. There are thousands of wayside images, as well as sculptures for shrines and graves, in a folk style characterized by shallow carving on a simple, coarse-stone shape. An outstanding type of folk painting flourished in the Ōtsu region from the 17th to the 19th century. Clearly distinct from the sophisticated ukiyo-e painting, it was executed by farmers and artisans and depicted folk as well as Buddhist deities, popular animal motifs such as the cock-and-hen, and popular characters and genre scenes, often satirical. There are also votive pictures, some portraying the horse and traceable to the ancient horse sacrifice. One of the late-surviving folk regions in modern Japan is on Sado island, where small cylindrical stone images are thrown into the sea to invoke pregnancy.

China. Chinese folk art must have been as extensive as any in the world, as evidenced by the descriptions of Western travellers and the souvenirs they collected and by various cultural and craft studies; but the problem of collating and analyzing the material as a folk category is forbidding. Every Chinese region has its own styles, and the entire art output is enormous. The art associated with weddings, funerals, and festivals is extravagant, even among the poor. In the country where paper was invented in antiquity, papermaking is a common skill, and the art of paper cutting is learned from childhood. Paper is used for the banner-like shop signs that give a special character

By courtesy of the Folkcraft Art Museum, Tokyo

"Cock and Hen," Ōtsu watercolour on paper, middle Edo period (1603–1867). In the Folkcraft Art Museum, Tokyo. 32.5 × 22.5 cm.



Balinese votive doll, dried palm leaves, c. 1880. In the Museum of Childhood, Edinburgh. Height 38 cm.

By courtesy of the Museum of Childhood Edinburgh

to Chinese streets and for many complicated models and festival objects.

Indonesia. In its effect on folk culture, the spread of Buddhism in the Far East has some parallels with the spread of Christianity in the West. In Indonesia, for example, where Buddhism penetrated an area whose local traditions were strong enough to survive and intermingle with the new concepts, there is much temple art of a folk character. Among the abundant ephemeral folk arts of Bali are the vegetal offerings and the beautifully stylized symbolic objects woven of palm leaf. Indonesian shadow puppets and printed textiles are world famous. (Ma.Ha.)

BIBLIOGRAPHY

Folklore: A. AARNE, *The Types of the Folktale,* 2nd rev., trans. and enl. by S. THOMPSON (1961), the standard index and reference work for the most widely distributed European folk narratives; F.J. CHILD (ed.), *The English and Scottish Popular Ballads,* 5 vol. (1882–98), the classic assemblage of 305 ballad types, with variant texts and learned headnotes; W.A. CLOUSTON, *Popular Tales and Fictions,* 2 vol. (1887), an early discussion of the wandering and migrations of folktales between India and Europe; M.R. COX, *Cinderella* (1893), the first comparative study of an international folktale, bringing together 345 variants, and contributing to the thesis of diffusion rather than independent invention of complex tales; L. DEGH, *Folktales and Society* (1969), a depth field study of storytelling behaviour in a Hungarian peasant community, with biographical portraits of leading folk narrators; R.M. DORSON, *American Folklore* (1959), an historical presentation of various folk traditions in the U.S., *The British Folklorists: A History,* and (ed.), *Peasant Customs and Savage Myths: Selections from the British Folklorists,* 2 vol. (both 1968), a history of the concept of folklore as it emerged in England with the 16th-century interest in antiquities and came to fruition among Victorian private scholars in the late 19th century—the volumes of selections present illustrative writings of the folklorists discussed in the history, and "Folktales of the World" (1963–), a series of authoritative volumes each of which is prepared by an eminent folktale scholar from the country represented; A.B. FRIEDMAN, *The Ballad Revival* (1961), interest in the ballad in England by antiquaries, poets, and the public treated in terms of literary history; JACOB GRIMM, *Teutonic Mythology,* trans. by J.S. STALLYBRASS, 4 vol. (1883–88), the encyclopaedic and influential work in which Grimm expounded his theory of the degeneration of an ancient high pantheon of Germanic deities into the extant fairy tales and witch beliefs of the contemporary peasantry; F.R.S. RAGLAN, *The Hero: A Study in Tradition, Myth, and Drama* (1956), a highly controversial explanation of all mythic narratives as following a uniform pattern derived from ancient sacrificial fertility rituals; Y.M. SOKOLOV, *Russian Folklore,* trans. by C.R. SMITH, (1950), a history of Russian folklore research given from the Soviet viewpoint regarding folklore as an expression of the class struggle; S. THOMPSON, *Motif-Index of Folk-Literature,* rev. ed., 6 vol. (1955–58), the major reference work in comparative folklore; J. VANSINA, *Oral Tradition* (1961), presentation of an historical methodology for the African historian enabling him to use oral historical chronicles and genealogies as legitimate source materials; D.K. WILGUS, *Anglo-American Folksong Scholarship Since 1898* (1959), a judicious appraisal of the recent schools of interpretation in ballad and folk-song studies.

Folk art in general: MAMIE HARMON, GIUSEPPE COCCHIARA, and ALESSANDRO MARABOTTINI MARABOTTI, "Folk Art," in the *Encyclopedia of World Art,* vol. 5, col. 452–483 (1961), present a broad, book-length theoretical survey focussed on visual art. A comprehensive work focussed on literature is MARIA LEACH (ed.), *Funk and Wagnalls Standard Dictionary of Folklore, Mythology and Legend* (1949). BRUNO NETTL, *Folk and Traditional Music of the Western Continents* (1956), offers much general theory in the course of treating his specific subject.

Folk literature: The best general treatment of the borderline between folk literature and sophisticated literature is H.M. and N. CHADWICK, *The Growth of Literature,* 3 vol. (1932–40). S. THOMPSON, *The Folktale* (1946), gives an introduction and extensive bibliography for the field of oral narrative literature. For myths of all parts of the world, see *Mythology of all Races,* 13 vol. (1916–32), valuable information, though some of the bibliographies are out of date. *The Journal of Folktale Studies* (3/yr.), and *FF Communications* (irreg.), are of primary importance. They include articles in English, French, and German. *FF Communications* is undoubtedly the leading series for all aspects of folklore. A good recent series of folktale collections in English is "Folktales of the World" ed. by R.M. DORSON. Important also is the much larger series in German, "Märchen der Weltliteratur." For the folktales of the ancient world good introductions are W.M.E. PETRIE (ed.), *Egyptian Tales Translated from the Papyri,* 2 vol. (1899); and S.N. KRAMER, *Sumerian Mythology: A Study of Spiritual and Literary Achievement in the 3d Millennium B.C.,* rev. ed. (1961). The standard Renaissance collection is *The Pentamerone of Giambattista Basile,* ed. and trans. from the Italian of BENEDETTO CROCE by N.M. PENZER, 2 vol. (1932). The commentaries on these tales are especially valuable. The new translation of Grimm's folktales, *German Folk Tales* by F.P. MAGOUN and A.H. KRAPPE (1960), is convenient for English readers. For folktales of the North American Indians, the American Negroes, and the peoples of Oceania and Israel, the following works are standard: W. MATTHEWS, *Navaho Legends* (1897); S. THOMPSON (ed.), *Tales of the North American Indians* (1929, reprinted 1966), an anthology with exhaustive comparative notes, valid until about 1926; M. JACOBS, *The Content and Style of an Oral Literature: Clackamas Chinook Myths and Tales* (1959), tales of a vanishing North Pacific tribe; W.A. LESSA, *Tales from Ulithi Atoll: A Comparative Study in Oceanic Folklore* (1961); K. LUOMALA, *Voices on the Wind. Polynesian Myths and Chants* (1955); R.M. DORSON, *American Negro Folktales* (1967); and D. NOY and D. BEN-AMOS (eds.), *Folktales of Israel* (1963). Types and classifications of folktales and legends are: A.A. AARNE, *The Types of the Folktale,* 2nd rev., trans. and enl. by S. THOMPSON (1961), a standard list of tales of old world provenance; S. THOMPSON, *Motif-Index of Folk Literature,* rev. ed., 6 vol. (1955–58); and R.T. CHRISTIANSEN, *The Migratory Legends* (1958), a classification of European legends. A good example of a survey of tales of a particular country is S. O'SUILLEABHAIN and R.T. CHRISTIANSEN, *Types of the Irish Folktale* (1963). A general introduction to folk song is G. HERZOG, "Song: Folksong, and the Music of Folksong," in *Funk and Wagnall's Standard Dictionary of Folklore,* vol. 2, pp. 1032–1050 (1949–50). A comprehensive introduction and listing of all the classical fables is in B.E. PERRY, *Babrius, and Phaedrus* (1965). The proverb, the riddle, and the charm are treated in A. TAYLOR, *The Proverb* (1962), *English Riddles from Oral Tradition* (1951); and W.R. BASCOM, *Ifa Divination: Communication Between Gods and Men in W. Africa* (1969). Important also is the work of L. DEGH, *Folktales and Society* (1969), a study of a Hungarian storyteller; A. DUNDES, *The Study of Folklore* (1965); V.A. PROPP, *Morphology of the Folktale* (1958); and T.A. SEBEOK, *Myth. A Symposium* (1965), a collection of theoretical treatments of mythology. An outstanding discussion of narrative folk song is A.B. LORD, *The Singer of Tales* (1960). A model historical and geographical study is W.E. ROBERTS, *The Tale of the Kind and the Unkind Girls* (1958).

Folk music: Among the scholarly periodicals devoted primarily to folk music, the most important are the *International Folk Music Council Yearbook* (1969–), formerly the *International Folk Music Council Journal; Ethnomusicology* (1953–); and the *Journal of American Folklore* (1888–). General works on folk music of Europe and the Americas are BRUNO NETTL, *Folk and Traditional Music of the Western Continents* (1965); WERNER DANCKERT, *Das europäische Volkslied* (1939); the lengthy and subdivided article on the folk music of many countries in *Grove's Dictionary of Music and Musicians,* 5th ed. (1955); and GEORGE HERZOG, "Song," in *Funk and Wagnalls Standard Dictionary of Folklore, Mythology and Legend* (1950). An im-

portant survey of the world's folk music in its relationship to certain characteristics of cultures is ALAN LOMAX, *Folk Song Style and Culture* (1968). WALTER WIORA, *Europäischer Volksgesang* (1952), provides an anthology of formal and melodic types in folk music; *Europäische Volksmusik und abendländische Tonkunst* (1957) explores the relationships between folk and classical music throughout European history. BELA BARTOK, *Hungarian Folk Music* (1931), is a classic study of one folk music style. A.B. LORD, *The Singer of Tales* (1960), deals with the epic traditions of eastern Europe. C.J. SHARP (comp.), *English Folk Songs from the Southern Appalachians*, 2 vol. (1932), is the pioneer collection of Anglo-American song; the total tune repertory of the most important traditional ballads in England and North America is published in B.H. BRONSON, *The Traditional Tunes of the Child Ballads*, 4 vol. (1959–70). The best general survey of Anglo-American folk song is R.D. ABRAHAMS and G. FOSS, *Anglo-American Folksong Style* (1968). The history of folk music research is treated in D.K. WILGUS, *Anglo-American Folksong Scholarship Since 1898* (1959). The modern urban folk song movement has given rise to a series of popular folk music periodicals, most of them ephemeral; notable American examples include *Broadside* (1962–), and *Sing Out!* (1950–).

Folk dance: V. ALFORD and R. GALLOP, *The Traditional Dance* (1935), authoritative, popular account of Europe's ancient ritual dances; C.M. BARBEAU (comp.), *Roundelays: Folk Dances and Games Collected in Canada and New Zealand* (1958), children's mimed rounds from French Canada, with descriptions, music, and bilingual texts; E. BURCHENAL, *Folk-Dances from Old Homelands* (1922), descriptions of European folk dances, for use in schools; N. CHILKOVSKY, *American Bandstand Dances in Labanotation* (1959), notations of jazz dances, for reconstruction by experts in the Laban system of notation; L.K. CZARNOWSKI, *Dances of Early California Days* (1950), splendid historical account, with descriptions and music, for use in schools; A.S. DUGGAN *et al.*, *The Folk Dance Library*, 5 vol. (1948), descriptions of dances from European and North American nations, with diagrams, music, and historical background, for school use; D.N. KENNEDY, *England's Dances* (1949), survey and interpretation of British ritual and folk dances; G.P. KURATH, *Iroquois Music and Dance*, U.S. Bureau of American Ethnology Bull. 187 (1964), and *Music and Dance of the Tewa Pueblos* (1970), analysis of dances and music, with many notation scores, interpretations, and background notes, not for reconstruction; LA MERI, *Spanish Dancing* (1948), skilled presentation of Spanish folk dances, with regional distinctions and some analysis of movement routines; J. LAWSON, *European Folk Dance* (1953), analysis of regional European steps and rhythms, with examples, useful facts, and questionable hypotheses; L. LEKIS, *Folk Dances of Latin America* (1958), exhaustive, annotated bibliography, with reliable comments on meanings and forms; M. MAYO, *American Square Dance*, rev. ed. (1948), a practical book for folk dance groups, with careful instructions and some music; C.J. SHARP (ed.), *The Country Dance Book*, 6 pt. (1909–22), exhaustive treatise on British folk dances by a scholarly pioneer, with diagrams and music; H.L. SPREEN, *Folk-Dances of South India* (1945), unusual, exotic material for schools, with movement descriptions, music, and bilingual texts; MARIA LEACH (ed.), *Dictionary of Folklore, Mythology and Legend*, vol. 1 (1949), definitions and scholarly interpretations.

Visual arts of the folk tradition: H.T. BOSSERT, *Ornamente der Volkskunst* (1949; Eng. trans., *Folk Art of Europe*, trans. by SYBIL MOHOLY NAGY, 1953, reprinted 1964), selection by the author from his *Volkskunst im Europa* (1926), a major compilation of folk designs, largely from textiles; D.P. BRANCH, *Folk Architecture of the East Mediterranean* (1966), includes Greek islands, central and southern Italy, with photos and diagrams; R.F. BUSSABARGER and B.D. ROBINS, *The Everyday Art of India* (1968), with glossary; ALFONSO CASO and D.F. RUBIN *et al.*, *Arte popular de México* (1963), a special issue of *Artes de México*, authoritative for crafts; E.O. CHRISTENSEN, *The Index of American Design* (1950), selections from a Federal Art Project study covering pre-1700–*c.* 1900; H.J. HANSEN (ed.), *Europas Volkskunst und die europäisch beeinflusste Volkskunst Amerikas* (1967; Eng. trans., *European Folk Art in Europe and the Americas*, 1968), country by country, chiefly European, with over 600 illustrations; M. HARMON *et al.*, "Folk Art," *Encyclopedia of World Art*, vol. 5, col. 451–506 (1961), a worldwide sampling of the arts, with extensive bibliography to *c.* 1960; STELLA KRAMRISCH, *Unknown India* (1968), an exhibition catalog of ritual and tribal folk art; FRANCES LICHTEN, *Folk Art of Rural Pennsylvania* (1946), German-American motifs and products; JEAN LIPMAN, *American Primitive Painting* (1942), pioneering study of folk painters; P.S. LORD and D.J. FOLEY, *The Folk Arts and Crafts of New England* (1965), over 500 illustrations of crafts; HUGO MUNSTERBERG, *The Folk Arts of Japan* (1958), includes the modern folk-art movement and living folk arts; BERNARD RUDOFSKY, *Architecture Without Architects* (1969), on primitive and vernacular styles all over the world; R.T. WILCOX, *Folk and Festival Costume of the World* (1965), over 150 regions, 111 plates, and bibliography.

Food Processing

Food processing is generally understood to encompass all methods by which raw foodstuffs are rendered suitable for cooking, consumption, or storage. This article treats the principal processing methods and products of the food-processing industry. Similar aspects of the beverage industry are treated in the article BEVERAGE PRODUCTION. For a discussion of the cultivation of food crops, see FARMING AND AGRICULTURAL TECHNOLOGY. For information on the nutritional role of various foods in the human diet, see NUTRITION. The culinary aspects of food preparation and consumption are treated in the article GASTRONOMY.

For more information on the coverage of topics related to food processing in both the *Macropædia* and *Micropædia,* see the *Propædia,* section 731, and the *Index.* (Ed.)

The article is divided into the following sections:

Food preservation and storage

Techniques for preserving food from natural deterioration, following harvest or slaughter, date to prehistoric times. Among the oldest methods of preservation are drying, refrigeration, and fermentation. Modern methods include canning, pasteurization, freezing, irradiation, and the addition of chemicals. Advances in packaging materials have played an important role in modern food preservation.

SPOILAGE MECHANISMS

Food spoilage may be defined as any change that renders food unfit for human consumption. These changes may be caused by various factors, including contamination by microorganisms, infestation by insects, or degradation by endogenous enzymes (those present naturally in the food). In addition, physical and chemical changes, such as the tearing of plant or animal tissues or the oxidation of certain constituents of food, may promote food spoilage. Foods obtained from plant or animal sources begin to spoil soon after harvest or slaughter. The enzymes contained in the cells of plant and animal tissues may be released as a result of any mechanical damage inflicted during postharvest handling. These enzymes begin to break down the cellular material. The chemical reactions catalyzed by the enzymes result in the degradation of food quality, such as the development of off-flavours, the deterioration of texture, and the loss of nutrients. The typical microorganisms that cause food spoilage are bacteria (*e.g., Lactobacillus*), yeasts (*e.g., Saccharomyces*), and molds (*e.g., Rhizopus*).

Microbial contamination. Bacteria and fungi (yeasts and molds) are the principal types of microorganisms that cause food spoilage and food-borne illnesses. Foods may be contaminated by microorganisms at any time during harvest, storage, processing, distribution, handling, or preparation. The primary sources of microbial contamination are soil, air, animal feed, animal hides and intestines, plant surfaces, sewage, and food processing machinery or utensils.

Bacteria. Bacteria are unicellular organisms that have a simple internal structure compared with the cells of other organisms. The increase in the number of bacteria in a population is commonly referred to as bacterial growth by microbiologists. This growth is the result of the division of one bacterial cell into two identical bacterial cells, a process called binary fission. Under optimal growth conditions, a bacterial cell may divide approximately every 20 minutes. Thus, a single cell can produce almost 70 billion cells in 12 hours. The factors that influence the growth of bacteria include nutrient availability, moisture, pH, oxygen levels, and the presence or absence of inhibiting substances (*e.g.,* antibiotics).

Growth conditions for bacteria

The nutritional requirements of most bacteria are chemical elements such as carbon, hydrogen, oxygen, nitrogen, phosphorus, sulfur, magnesium, potassium, sodium, calcium, and iron. The bacteria obtain these elements by utilizing gases in the atmosphere and by metabolizing certain food constituents such as carbohydrates and proteins.

Temperature and pH play a significant role in controlling the growth rates of bacteria. Bacteria may be classified into three groups based on their temperature requirement for optimal growth: thermophiles (55°–75° C, or 130°–170° F), mesophiles (20°–45° C, or 70°–115° F), or psychrotrophs (10°–20° C, or 50°–70° F). In addition, most bacteria grow best in a neutral environment (pH equal to 7).

Bacteria also require a certain amount of available water for their growth. The availability of water is expressed as water activity and is defined by the ratio of the vapour pressure of water in the food to the vapour pressure of pure water at a specific temperature. Therefore, the water activity of any food product is always a value between 0 and 1, with 0 representing an absence of water and 1 representing pure water. Most bacteria do not grow in foods with a water activity below 0.91, although some halophilic bacteria (those able to tolerate high salt concentrations) can grow in foods with a water activity lower than 0.75. Growth may be controlled by lowering the water activity—either by adding solutes such as sugar, glycerol, and salt or by removing water through dehydration.

The oxygen requirements for optimal growth vary considerably for different bacteria. Some bacteria require the presence of free oxygen for growth and are called obligate aerobes, whereas other bacteria are poisoned by the presence of oxygen and are called obligate anaerobes. Facultative anaerobes are bacteria that can grow in both the presence or absence of oxygen. In addition to oxygen concentration, the oxygen reduction potential of the growth medium influences bacterial growth. The oxygen reduction potential is a relative measure of the oxidizing or reducing capacity of the growth medium.

When bacteria contaminate a food substrate, it takes some time before they start growing. This lag phase is the period when the bacteria are adjusting to the environment. Following the lag phase is the log phase, in which population grows in a logarithmic fashion. As the population grows, the bacteria consume available nutrients and produce waste products. When the nutrient supply is depleted, the growth rate enters a stationary phase in which the number of viable bacteria cells remains the same. During the stationary phase, the rate of bacterial cell growth is equal to the rate of bacterial cell death. When the rate of cell death becomes greater than the rate of cell growth, the population enters the decline phase.

Table 1: Enzymes That Cause Food Spoilage

enzyme	food product	spoilage action
Ascorbic acid oxidase	vegetables	destruction of vitamin C
Lipase	cereals	discoloration
	milk	hydrolytic rancidity
	edible oils	hydrolytic rancidity
Lipoxygenase	vegetables	destruction of vitamin A, off-flavour
Pectic enzyme	citrus juices	destruction of pectic substances
	fruits	excessive softening
Peroxidase	fruits	browning
Polyphenoloxidase	fruits, vegetables	browning, off-flavour, vitamin loss
Protease	eggs	reduce shelf life of fresh and dried whole eggs
	crab, lobster	overtenderization
	flour	reduction of gluten formation
Thiaminase	meats, fish	destruction of thiamine

A bacterial population is expressed either per gram or per square centimetre of surface area. Rarely does the total bacterial population exceed 10^{10} cells per gram. A population of less than 10^6 cells per gram does not cause any noticeable spoilage except in raw milk. Populations of between 10^6 and 10^7 cells per gram cause spoilage in some foods; for example, they can generate off-odours in vacuum-packaged meats. Populations of between 10^7 and 10^8 cells per gram produce off-odours in meats and some vegetables. At levels above 5×10^7 cells per gram, most foods exhibit some form of spoilage.

When the conditions for bacterial cell growth are unfavourable (*e.g.,* low or high temperatures or low moisture content), several species of bacteria can produce resistant cells called endospores. Endospores are highly resistant to heat, chemicals, desiccation (drying out), and ultraviolet light. The endospores may remain dormant for long periods of time. When conditions become favourable for growth (*e.g.,* thawing of meats), the endospores germinate and produce viable cells that can begin exponential growth.

Fungi. The two types of fungi that are important in food spoilage are yeasts and molds. Molds are multicellular fungi that reproduce by the formation of spores (single cells that can grow into a mature fungus). Spores are formed in large numbers and are easily dispersed through the air. Once these spores land on a food substrate, they can grow and reproduce if conditions are favourable. Yeasts are unicellular fungi that are much larger than bacterial cells. They reproduce by cell division (binary fission) or budding.

Growth conditions for fungi

The conditions affecting the growth of fungi are similar to those affecting bacteria. Both yeasts and molds are able to grow in an acidic environment (pH less than 7). The pH range for yeast growth is 3.5 to 4.5 and for molds is 3.5 to 8.0. The low pH of fruits is generally unfavourable for the growth of bacteria, but yeasts and molds can grow and cause spoilage in fruits. For example, species of the fungal genus *Colletotrichum* cause crown rot in bananas. Yeasts promote fermentation in fruits by breaking down sugars into alcohol and carbon dioxide. The amount of available water in a food product is also critical for the growth of fungi. Yeasts are unable to grow at a water activity of less than 0.9, and molds are unable to grow at a water activity below 0.8.

Control of microbial contamination. The most common methods used either to kill or to reduce the growth of microorganisms are the application of heat, the removal of water, the lowering of temperature during storage, the reduction of pH, the control of oxygen and carbon dioxide concentrations, and the removal of the nutrients needed for growth. The use of chemicals as preservatives is strictly regulated by governmental agencies such as the Food and Drug Administration (FDA) in the United States. Although a chemical may have preservative functions, its safety must be proved before it may be used in food products. To suppress yeast and mold growth in foods, a number of chemical preservatives are permitted. In the United States, the list of such chemicals, known as GRAS (Generally Recognized as Safe), includes compounds such as benzoic acid, sodium benzoate, propionic acid, sorbic acid, and sodium diacetate.

Chemical deterioration. *Enzymatic reactions.* Enzymes are large protein molecules that act as biological catalysts, accelerating chemical reactions without being consumed to any appreciable extent themselves. The activity of enzymes is specific for a certain set of chemical substrates, and it is dependent on both pH and temperature.

The living tissues of plants and animals maintain a balance of enzymatic activity. This balance is disrupted upon harvest or slaughter. In some cases, enzymes that play a useful role in living tissues may catalyze spoilage reactions following harvest or slaughter. For example, the enzyme pepsin is found in the stomach of all animals and is involved in the breakdown of proteins during the normal digestion process. However, soon after the slaughter of an animal, pepsin begins to break down the proteins of the organs, weakening the tissues and making them more susceptible to microbial contamination. After the harvesting of fruits, certain enzymes remain active within the cells of the plant tissues. These enzymes continue to catalyze the biochemical processes of ripening and may eventually lead to rotting, as can be observed in bananas. In addition, oxidative enzymes in fruits continue to carry out cellular respiration (the process of using oxygen to metabolize glucose for energy). This continued respiration decreases the shelf life of fresh fruits and may lead to spoilage. Respiration may be controlled by refrigerated storage or modified-atmosphere packaging. Table 1 lists a number of enzymes involved in the degradation of food quality.

Autoxidation. The unsaturated fatty acids present in the lipids of many foods are susceptible to chemical breakdown when exposed to oxygen. The oxidation of unsaturated fatty acids is autocatalytic; that is, it proceeds by a free-radical chain reaction. Free radicals contain an unpaired electron (represented by a dot in the molecular formula) and, therefore, are highly reactive chemical molecules. The basic mechanisms in a free-radical chain reaction involve initiation, propagation, and termination steps (Figure 1). Under certain conditions, in initiation a free-radical molecule ($X\cdot$) present in the food removes a hydrogen (H) atom from a lipid molecule, producing a lipid radical ($L\cdot$). This lipid radical reacts with molecular oxygen (O_2) to form a peroxy radical ($LOO\cdot$). The peroxy radical removes a hydrogen atom from another lipid molecule and the reaction starts over again (propagation). During the propagation steps, hydroperoxide molecules (LOOH) are formed that may break down into alkoxy ($LO\cdot$) and peroxy radicals plus water (H_2O). The lipid, alkoxy, and peroxy radicals may combine with one another (or other radicals) to form stable, nonpropagating products (termination). These products result in the development of rancid off-flavours. In addition to promoting rancidity, the free radicals and peroxides produced in these reactions may have other negative effects, such as the bleaching of food colour and the destruction of vitamins A, C, and E. This type of deterioration is prevalent in fried snacks, nuts, cooking oils, and margarine.

Rancidity

Initiation: $X\cdot + LH \rightarrow L\cdot + XH$

Propagation:
$$L\cdot + O_2 \rightarrow LOO\cdot$$
$$LOO\cdot + LH \rightarrow LOOH + L\cdot$$
$$2LOOH \rightarrow LO\cdot + LOO\cdot + H_2O$$

Termination: $L\cdot$, $LO\cdot$, $LOO\cdot$ → a number of stable nonpropagating species

Figure 1: The autoxidation of unsaturated fatty acids.

Maillard reaction. Another chemical reaction that causes major food spoilage is nonenzymatic browning, also known as the Maillard reaction. This reaction takes place between reducing sugars (simple monosaccharides capable of carrying out reduction reactions) and the amino group of proteins or amino acids present in foods. The products of the Maillard reaction lead to a darkening of colour, reduced solubility of proteins, development of bitter flavours, and reduced nutritional availability of certain amino acids such as lysine. The rate of this reaction is influenced by the water activity, temperature, and pH of the food product. Nonenzymatic browning causes spoilage during the storage of dry milk, dry whole eggs, and breakfast cereals.

Light-induced reactions. Light influences a number of chemical reactions that lead to spoilage of foods. These light-induced reactions include the destruction of chlorophyll (the photosynthetic pigment that gives plants their green colour), resulting in the bleaching of certain vegetables; the discoloration of fresh meats; the destruction of riboflavin in milk; and the oxidation of vitamin C and carotenoid pigments (a process called photosensitized oxidation). The use of packaging material that prevents exposure to light is one of the most effective means of preventing light-induced chemical spoilage.

LOW-TEMPERATURE PRESERVATION

Storage at low temperatures prolongs the shelf life of many foods. In general, low temperatures reduce the growth rates of microorganisms and slow many of the physical and chemical reactions that occur in foods.

Refrigeration. The life of many foods may be increased by storage at temperatures below 4° C (40° F). Commonly refrigerated foods include fresh fruits and vegetables, eggs, dairy products, and meats. Some foods, such as tropical fruits (*e.g.,* bananas), are damaged if exposed to low temperatures. Also, refrigeration cannot improve the quality of decayed food; it can only retard deterioration. One problem of modern mechanical refrigeration—that of dehydration of foods due to moisture condensation—has been overcome through humidity control mechanisms within the storage chamber and by appropriate packaging techniques.

Freezing. Freezing and frozen storage provide an excellent means of preserving the nutritional quality of foods. At subfreezing temperatures the nutrient loss is extremely slow for the typical storage period used in commercial trade.

History. Early freezing methods were based on the principle that mixing salt with ice results in temperatures well below 0° C (32° F). By the end of the 19th century, this method was being used commercially in the United States to freeze fish and poultry. By the 1920s Clarence Birdseye had developed two processes for freezing fish based on his quick freezing theory. His first patent, describing a method for preserving piscatorial products, involved placing food between two metal plates that were chilled by a calcium chloride solution to approximately −40° C (−40° F). The second process utilized two hollow metal plates that were cooled to −25° C (−13° F) by vaporization of ammonia. This freezing apparatus was the forerunner of the multiple plate freezer that is widely used in the modern food industry.

Crystal growth

The freezing process. The freezing of food involves lowering its temperature below 0° C, resulting in the gradual conversion of water, present in the food, into ice. Freezing is a crystallization process that begins with a nucleus or a seed derived from either a nonaqueous particle or a cluster of water molecules (formed when the temperature is reduced below 0° C). This seed must be of a certain size to provide an adequate site for the crystal to begin to grow. If physical conditions are conducive to the presence of numerous seeds for crystallization, then a large number of small ice crystals will form. However, if only a few seeds are initially available, then a few ice crystals will form and each will grow to a large size. The size and the number of ice crystals influence the final quality of many frozen foods; for example, the smooth texture of ice cream indicates the presence of a large number of small ice crystals.

In pure water, the freezing process is initiated by lowering the temperature to slightly below 0° C, called supercooling. As ice crystals begin to grow, the temperature returns to the freezing point. During the conversion of liquid water to ice, the temperature of the system does not change. The heat removed during this step is called the latent heat of fusion (equivalent to 333 joules per gram of water). Once all the water is converted to ice, any additional removal of heat will result in a decrease in the temperature below 0° C.

The freezing of foods exhibits a number of important differences from the freezing of pure water. Foods do not freeze at 0° C. Instead, owing to the presence of different soluble particulates (solutes) in the water present in foods, most foods begin to freeze at a temperature between 0° and −5° C (32° and 23° F). In addition, the removal of latent heat in foods during freezing does not occur at a fixed temperature. As the water present in the food freezes into ice, the remaining water becomes more concentrated with solutes. As a result, the freezing point is further depressed. Therefore, foods have a zone of maximum ice crystal formation that typically extends from −1° to −4° C (30° to 25° F). Damage to food quality during freezing can be minimized if the temperature of the product is brought below this temperature range as quickly as possible.

Industrial freezers. The rate at which heat is removed from a food during freezing depends on how fast heat can travel within the food and how efficiently it can be liberated from the surface of the food into the surrounding atmosphere. Industrial freezers remove heat from the surface of a food as rapidly as possible. There are several types of industrial freezers, including air-blast tunnel freezers, belt freezers, fluidized-bed freezers, plate freezers, and cryogenic freezers.

Air-blast freezers

In air-blast tunnel freezers and belt freezers, precooled air at approximately −40° C is blown over the food products. Packaged foods, such as fruits, vegetables, bakery goods, poultry, meats, and prepared meals, are usually frozen in air-blast tunnels. The packages are placed onto dollies or hand trucks and then rolled into the freezer tunnels. In a belt freezer, food is placed on a conveyor belt that moves through a freezing zone. Bakery goods, chicken parts, and meat patties are frozen using a belt freezer.

Fluidized-bed freezers are used to freeze particulate foods such as peas, cut corn, diced carrots, and strawberries. The foods are placed on a mesh conveyor belt and moved through a freezing zone in which cold air is directed upward through the mesh belt and the food particulates begin to tumble and float. This tumbling exposes all sides of the food to the cold air and minimizes the resistance to heat transfer at the surface of the food.

Plate freezers are used to freeze flat products, such as pastries, fish fillets, and beef patties, as well as irregular-shaped vegetables that are packaged in brick-shaped containers, such as asparagus, cauliflower, spinach, and broccoli. The food is firmly pressed betwcen metal plates that are cooled to subfreezing temperatures by internally circulating refrigerants.

Cryogenic freezing is used to freeze food at an extremely fast rate. The food is moved through a spray of liquid nitrogen or directly immersed in liquid nitrogen. The liquid nitrogen boils around the food at a temperature of −196° C (−321° F) and extracts a large amount of heat.

Quality of frozen foods. Improper freezing or storage of foods may result in detrimental quality changes. When foods with high amounts of water are frozen slowly, they may experience a loss of fluid, called drip, upon thawing. This fluid loss causes dehydration and nutrient loss in frozen food products.

During frozen storage, the ice crystals present in foods may enlarge in size, producing undesirable changes in texture. This phenomenon is commonly observed when the storage temperature is allowed to fluctuate. For example, ice cream stored in an automatic defrosting domestic freezer becomes sandy in texture because the ice crystals increase in size as the temperature of the system fluctuates.

Improperly packaged frozen foods lose small amounts of moisture during storage, resulting in surface dehydration (commonly called freezer burn). Frozen meats with freezer burn have the appearance of brown paper and quickly become rancid. Freezer burn can be minimized by the use of tightly wrapped packages and the elimination of fluctuating temperatures during storage.

© Mark E. Gibson

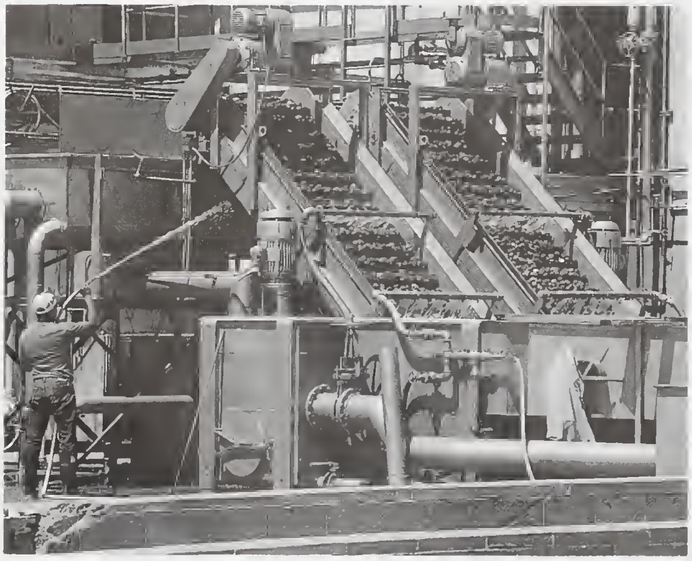


Figure 2: Spray washing of harvested tomatoes prior to processing.

THERMAL PROCESSING

Thermal processing is defined as the combination of temperature and time required to eliminate a desired number of microorganisms from a food product.

Canning. Nicolas Appert, a Parisian confectioner by trade, is credited with establishing the heat processing of foods as an industry. In 1810 he received official recognition for his process of enclosing food in bottles, corking the bottles, and placing the bottles in boiling water for various periods of time. In the same year Peter Durand received a British patent for the use of containers made of glass, pottery, tin, or other metals for the heat preservation of foods. In 1822 Ezra Daggett and Thomas Kensett announced the availability of preserved foods in tin cans in the United States. Tin-coated steel containers, made from 98.5 percent sheet steel with a thin coating of tin, soon became common. These cans had a double seamed top and bottom to provided an airtight seal and could be manufactured at high speeds.

The establishment of the canning process on a more scientific basis did not occur until 1896, when the microorganism *Clostridium botulinum,* with its lethal toxin causing botulism, was discovered by Émile van Ermengem.

Presterilization procedures. Selected crop varieties are grown specially for canning purposes. The harvesting schedules of the crops are carefully selected to conform to the cannery operations. A typical canning operation involves cleaning, filling, exhausting, can sealing, heat processing, cooking, labeling, casing, and storage. Most of these operations are performed using high-speed, automatic machines.

Cleaning involves the use of shakers, rotary reel cleaners, air blasters, water sprayers (as shown in Figure 2), or immersion washers. Any inedible or extraneous material is removed before washing, and only potable water is used in the cleaning systems.

Evacuation of air

Automatic filling machines are used to place the cleaned food into cans or other containers, such as glass jars or plastic pouches. When foods containing trapped air, such as leafy vegetables, are canned, the air must be removed from the cans prior to closing and sealing the lids by a process called exhausting. Exhausting is accomplished using steam exhaust hoods or by creation of a vacuum.

Immediately after exhausting, the lids are placed on the cans and the cans are sealed. An airtight seal is achieved between the lid and the rim of the can using a thin layer of gasket or compound. The anaerobic conditions prevent the growth of oxygen-requiring microorganisms. In addition, many of the spores of anaerobic microorganisms are less resistant to heat and are easily destroyed during the heat treatment.

Sterilization. The time and temperature required for the sterilization of foods are influenced by several factors, including the type of microorganisms found on the food, the size of the container, the acidity or pH of the food, and the method of heating.

The thermal processes of canning are generally designed to destroy the spores of the bacterium *C. botulinum.* This microorganism can easily grow under anaerobic conditions, producing the deadly toxin that causes botulism. Sterilization requires heating to temperatures greater than 100° C (212° F). However, *C. botulinum* is not viable in acidic foods that have a pH less than 4.6. These foods can be adequately processed by immersion in water at temperatures just below 100° C.

The sterilization of low-acid foods (pH greater than 4.6) is generally carried out in steam vessels called retorts at temperatures ranging from 116° to 129° C (240° to 265° F). The retorts are controlled by automatic devices, and detailed records are kept of the time and temperature treatments for each lot of processed cans. At the end of the heating cycle, the cans are cooled under water sprays or in water baths to approximately 38° C (100° F) and dried to prevent any surface rusting. The cans are then labeled, placed in fibreboard cases either by hand or machine, and stored in cool, dry warehouses.

Quality of canned foods. The sterilization process is designed to provide the required heat treatment to the slowest heating location inside the can, called the cold spot. The areas of food farthest from the cold spot get a more severe heat treatment that may result in overprocessing and impairment of the overall quality of the product. Flat, laminated pouches can reduce the heat damage caused by overprocessing.

A significant loss of nutrients, especially heat-labile vitamins, may occur during the canning process. In general, canning has no major effect on the carbohydrate, protein, or fat content of foods. Vitamins A and D and beta-carotene are resistant to the effects of heat. However, vitamin B_1 is sensitive to thermal treatment and the pH of the food. Although the anaerobic conditions of canned foods have a protective effect on the stability of vitamin C, it is destroyed during long heat treatments.

The ends of processed cans are slightly concave because of the internal vacuum created during sealing. Any bulging of the ends of a can may indicate a deterioration in quality due to mechanical, chemical, or physical factors. This bulging may lead to swelling and possible explosion of the can.

Pasteurization. Pasteurization is the application of heat to a food product in order to destroy pathogenic (disease-producing) microorganisms, to inactivate spoilage-causing enzymes, and to reduce or destroy spoilage microorganisms. The relatively mild heat treatment used in the pasteurization process causes minimal changes in the sensory and nutritional characteristics of foods compared to the severe heat treatments used in the sterilization process.

Pasteurizing acidic and basic foods

The temperature and time requirements of the pasteurization process are influenced by the pH of the food. When the pH is below 4.5, spoilage microorganisms and enzymes are the main targets of pasteurization. For example, the pasteurization process for fruit juices is aimed at inactivating certain enzymes such as pectinesterase and polygalacturonase. The typical processing conditions for the pasteurization of fruit juices include heating to 77° C (171° F) and holding for 1 minute, followed by rapid cooling to 7° C (45° F). In addition to inactivating enzymes, these conditions destroy any yeasts or molds that may lead to spoilage. Equivalent conditions capable of reducing spoilage microorganisms involve heating to 65° C (149° F) and holding for 30 minutes or heating to 88° C (190° F) and holding for 15 seconds.

When the pH of a food is greater than 4.5, the heat treatment must be severe enough to destroy pathogenic bacteria. In the pasteurization of milk, the time and temperature conditions target the pathogenic bacteria *Mycobacterium tuberculosis, Coxiella burnetti,* and *Brucella abortus.* The typical heat treatment used for pasteurizing milk is 72° C (162° F) for 15 seconds, followed by rapid cooling to 7° C. Other equivalent heat treatments include heating to 63° C (145° F) for 30 minutes, 90° C (194° F) for 0.5 second, and 94° C (201° F) for 0.1 second. The high-temperature-short-time (HTST) treatments cause less damage to the nutrient composition and sensory characteristics of foods and therefore are preferred over the low-temperature-long-time (LTLT) treatments.

Since the heat treatment of pasteurization is not severe enough to render a product sterile, additional methods such as refrigeration, fermentation, or the addition of chemicals are often used to control microbial growth and to extend the shelf life of a product. For example, the pasteurization of milk does not kill thermoduric bacteria (those resistant to heat), such as *Lactobacillus* and *Streptococcus,* or thermophilic bacteria (those that grow at high temperatures), such as *Bacillus* and *Clostridium.* Therefore, pasteurized milk must be kept under refrigerated conditions.

Liquid foods such as milk, fruit juices, beers, wines, and liquid eggs are pasteurized using plate-type heat exchangers. Wine and fruit juices are normally deaerated prior to pasteurization in order to remove oxygen and minimize oxidative deterioration of the products. Plate-type heat exchangers consist of a large number of thin, vertical steel plates that are clamped together in a frame. The plates are separated by small gaskets that allow the liquid to flow between each successive plate. The liquid product and heating medium (*e.g.,* hot water) are pumped through alternate channels, and the gaskets ensure that the liquid product and heating or cooling mediums are kept separate. Plate-type heat exchangers are effective in rapid heating and cooling applications. After the pasteurization process is completed, the product is packaged under aseptic conditions to prevent recontamination of the product.

Aseptic processing. The aseptic process involves placing a sterilized product into a sterilized package that is then sealed under sterile conditions. It began in 1914 with the development of sterile filters for use in the wine industry. However, because of unreliable machinery, it remained commercially unsuccessful until 1948 when William McKinley Martin helped develop the Martin system, which later became known as the Dole Aseptic Canning System. This system involved the sterilization of liquid foods by rapidly heating them in tubular heat exchangers, followed by holding and cooling steps. The cans and lids were sterilized with superheated steam, and the sterilized containers were filled with the sterile liquid food. The lids were then sealed in an atmosphere of superheated steam. By the 1980s hydrogen peroxide was being used throughout Europe and the United States for the sterilization of polyethylene surfaces.

Commercial sterility. In aseptic processing the thermal process is based on achieving commercial sterility—*i.e.,* no more than 1 nonsterile package for every 10,000 processed packages. The aseptic process uses the high-temperature-short-time (HTST) method in which foods are heated at a high temperature for a short period of time. The time and temperature conditions depend on several factors, such as size, shape, and type of food. The HTST method results in a higher retention of quality characteristics, such as vitamins, odour, flavour, and texture, while achieving the same level of sterility as the traditional canning process in which food is heated at a lower temperature for a longer period of time.

The heating and cooling of liquid foods can be performed using metal plate heat exchangers. These heat exchangers have large surface areas that result in improved heating and cooling rates. Other types of heat exchangers involve surrounding the food with steam or directly injecting steam into the food. Products sterilized with steam are then pumped into a vacuum chamber, where they are cooled rapidly.

Liquid foods that contain large solid particles are heated in scraped-surface heat exchangers. These heat exchangers use blades to continuously scrape the inside surface of the heating chamber. The scraping action protects highly viscous foods from being burned on the heating surface.

Ohmic heating

An alternate method for heating foods, called ohmic heating, passes a low-frequency electric current of 50 to 60 hertz directly through the food. A liquid food containing solids, such as diced fruit, is pumped through a pipe surrounded by electrodes. The product is heated as long as the electrical conductivity of the food is uniform throughout the entire volume. This uniform rate of heating prevents the overprocessing of any individual region of the food. Ohmic heating yields a food product of higher quality than those processed using conventional systems.

Packaging aseptically processed products. The packaging containers used in aseptic processing are sterilized separately before they are used. The packaging machinery is sterilized using steam, sterile gases, or hydrogen peroxide. The sterilization process is generally monitored by culturing a test organism. For example, the remaining presence of the highly heat-resistant bacterium *Bacillus subtilis globigii* can be used as a marker to measure the completeness of sterilization.

Packages must be sealed under sterile conditions, usually using high-temperature sealing plates. Foods that are aseptically processed do not require refrigeration for storage.

Blanching. Blanching is a thermal process used mostly for vegetable tissues prior to freezing, drying, or canning. Before canning, blanching serves several purposes, including cleaning of the product, reducing the microbial load, removing any entrapped gases, and wilting the tissues of leafy vegetables so that they can be easily put into the containers. Blanching also inactivates enzymes that cause deterioration of foods during frozen storage.

Blanching is carried out at temperatures close to 100° C (212° F) for two to five minutes in either a water bath or a steam chamber. Because steam blanchers use a minimal amount of water, extra care must be taken to ensure that the product is uniformly exposed to the steam. Steam blanching leafy vegetables is especially difficult because they tend to clump together. The effectiveness of the blanching treatment is usually determined by measuring the residual activity of an enzyme called peroxidase.

CONTROLLING WATER ACTIVITY

Foods containing high concentrations of water are generally more susceptible to deterioration by microbial contamination and enzymatic activity. The water content of foods can be controlled by removing water through dehydration or by adding solutes to the food. In both cases the concentration of solutes in the food increases and the concentration of water decreases.

Dehydration. Dehydration, or drying, of foods has long been practiced commercially in the production of spaghetti and other starch products. As a result of advances made during World War II, the technique has been applied to a growing list of food products, including fruits, vegetables, skim milk, potatoes, soup mixes, and meats.

Pathogenic (toxin-producing) bacteria occasionally withstand the unfavourable environment of dried foods, causing food poisoning when the product is rehydrated and eaten. Control of bacterial contaminants in dried foods requires high-quality raw materials having low contamination, adequate sanitation in the processing plant, pasteurization before drying, and storage conditions that protect from infection by dust, insects, and rodents or other animals.

Foodstuffs may be dried in air, superheated steam, vacuum, or inert gas or by direct application of heat. Air is the most generally used drying medium, because it is plentiful and convenient and permits gradual drying, allowing sufficient control to avoid overheating that might result in scorching and discoloration. Air may be used both to transport heat to the food being dried and to carry away liberated moisture vapour. The use of other gases requires special moisture recovery systems.

Effect of drying on nutrients

Loss of moisture content produced by drying results in increased concentration of nutrients in the remaining food mass. The proteins, fats, and carbohydrates in dried foods are present in larger amounts per unit weight than in their fresh counterparts, and the nutrient value of most reconstituted or rehydrated foods is comparable to that of fresh items. The biological value of dried protein is dependent, however, on the method of drying. Prolonged exposure to high temperatures can render the protein less useful in the diet. Low-temperature treatment, on the other hand, may increase the digestibility of protein. Some vitamins are sensitive to the dehydration process. For example, in dried meats significant amounts of vitamin C and the B vitamins—riboflavin, thiamine, and niacin—are lost during dehydration.

Dried eggs, meat, milk, and vegetables are ordinarily packaged in tin or aluminum containers. Fibreboard or other types of material may be employed but are less satisfactory than metal, which offers protection against insects and moisture loss or gain and which permits packaging with an inert gas.

In-package desiccants (drying agents) improve storage stability of dchydrated white potatoes, sweet potatoes, cabbage, carrots, bcets, and onions and give substantial protection against browning. Retention of ascorbic acid (vitamin C) is markedly improved by packaging at temperatures up to 49° C (120° F); the packaging gas may be either nitrogcn or air.

Freeze-drying process

A related technique, freezc-drying, employs high vacuum conditions, pcrmitting establishment of specific tcmperaturc and prcssure conditions. The raw food is frozen, and thc low pressurc conditions causc the ice in the food to sublimate directly into vapour (*i.e.,* it does not transit through the liquid state). Adequate control of processing conditions contributes to satisfactory rehydration, with substantial retention of nutrient, colour, flavour, and texture characteristics.

Concentration of moist foods. Foods with substantial acidity, when concentrated to 65 percent or more soluble solids, may be preserved by mild heat treatments. High acid content is not a requirement for preserving foods concentrated to over 70 percent solids.

Fruit jelly and preserve manufacture, an important fruit by-product industry, is based on the high-solids–high-acid principle, with its moderate heat-treatment requirements. Fruits that possess excellent qualities but are visually unattractive may be preserved and utilized in the form of concentrates, which have a pleasing taste and substantial nutritive value.

Jellies and other fruit preserves are prepared from fruit by adding sugar and concentrating by evaporation to a point where microbial spoilage cannot occur. The prepared product can be stored without hermetic sealing, although such protection is useful to control mold growth, moisture loss, and oxidation. In modern practice, vacuum sealing has replaced the use of a paraffin cover.

The jelly-forming characteristics of fruits and their extracts are due to pectin, a substance present in varying amounts in all fruits. The essential ingredients in a fruit gel are pectin, acid, sugar, and water. Flavouring and colouring agents may be added, and additional pectin and acid may be added to overcome any deficiencies in the fruit itself.

Candied and glacéed fruits are made by slow impregnation of the fruit with syrup until the concentration of sugar in the tissue is sufficiently high to prevent growth of spoilage microorganisms. The candying process is conducted by treating fruits with syrups of progressively increasing sugar concentrations, so that the fruit does not soften into jam or become tough and leathery. After sugar impregnation the fruit is washed and dried. The resulting candied fruit may be packaged and marketed in this condition or may be dipped into syrup, becoming coated with a thin glazing of sugar (glacéed) and again dried.

FERMENTATION AND PICKLING

Although microorganisms are usually thought of as causing spoilage, they are capable under certain conditions of producing desirable effects, including oxidative and alcoholic fermentation. The microorganisms that grow in a food product, and the changes they produce, are determined by acidity, available carbohydrates, oxygen, and temperature. An important food preservation method combines salting to control microorganisms selectively and fermentation to stabilize the treated tissues.

Role of salt

Pickled fruits and vegetables. Fresh fruits and vegetables soften after 24 hours in a watery solution and begin a slow, mixed fermentation-putrefaction. The addition of salt suppresses undesirable microbial activity, creating a favourable environment for the desired fermentation. Most green vegetables and fruit may be preserved by pickling.

When the pickling process is applied to a cucumber, its fermentable carbohydrate reserve is turned into acid, its colour changes from bright green to olive or yellow-green, and its tissue becomes translucent. The salt concentration is maintained at 8 to 10 percent during the first week and is increased 1 percent a week thereafter until the solution reaches 16 percent. Under properly controlled conditions the salted, fermented cucumber, called salt stock, may be held for several years.

Salt stock is not a consumer commodity. It must be freshened and prepared into consumer items. In cucumbers this is accomplished by leaching the salt from the cured cucumber with warm water (43°–54° C [110°–130° F]) for 10 to 14 hours. This process is repeated at least twice, and, in the final wash, alum may be added to firm the tissue and turmeric to improve the colour.

Pickled meat. Meat may be preserved by dry curing or with a pickling solution. The ingredients used in curing and pickling are sodium nitrate, sodium nitrite, sodium chloride, sugar, and citric acid or vinegar.

Various methods are used: the meat may be mixed with dry ingredients; it may be soaked in pickling solution; pickling solution may be pumped or injected into the flesh; or a combination of these methods may be used.

Curing may be combined with smoking. Smoke acts as a dehydrating agent and coats the meat surfaces with various chemicals, including small amounts of formaldehyde.

Deterioration of fermented and pickled products. Fermented foods and pickled products require protection against molds, which metabolize the acid developed and allow the advance of other microorganisms. Fermented and pickled food products placed in cool storage can be expected to remain stable for several months. Longer storage periods demand more complete protection, such as canning.

Nutrient retention in fermented and pickled products is about equal to retention for products preserved by other methods. Carbohydrates usually undergo conversion to acid or to alcohol, but these are also of nutritive value. In some instances, nutrient levels are increased because of the presence of yeasts.

CHEMICAL PRESERVATION

Chemical food preservatives are substances which, under certain conditions, either delay the growth of microorganisms without necessarily destroying them or prevent deterioration of quality during manufacture and distribution. The former group includes some natural food constituents which, when added to foods, retard or prevent the growth of microorganisms. Sugar is used partly for this purpose in making jams, jellies, and marmalades and in candying fruit. The use of vinegar and salt in pickling and of alcohol in brandying also falls in this category. Some chemicals foreign to foods are added to prevent the growth of microorganisms. The latter group includes some natural food constituents such as ascorbic acid (vitamin C), which is added to frozen peaches to prevent browning, and a long list of chemical compounds foreign to foods and classified as antioxidants, bleaching agents, acidulants, neutralizers, stabilizers, firming agents, and humectants.

Benzoates

Organic chemical preservatives. Sodium benzoate and other benzoates are among the principal chemical preservatives. The use of benzoates in certain products in prescribed quantity (usually not exceeding 0.1 percent) is permitted in most countries, some of which require a declaration of its use on the label of the food container. Since free benzoic acid actually is the active agent, benzoates must be used in an acid medium in order to be effective. The ability of cranberries to resist rapid deterioration is attributed to their high benzoic acid content. Benzoic acid is more effective against yeasts than against molds and bacteria.

Other organic compounds used as preservatives include vanillic acid esters, monochloroacetic acid, propionates, sorbic acid, dehydroacetic acid, and glycols.

Inorganic chemical preservatives. Sulfur dioxide and sulfites are perhaps the most important inorganic chemical preservatives. Sulfites are more effective against molds than against yeasts and are widely used in the preservation of fruits and vegetables. Sulfur compounds are extensively used in wine making and, as in most other instances when this preservative is used, much care has to be exercised to keep the concentrations low in order to avoid undesirable effects on flavour.

Oxidizing agents such as nitrates and nitrites are commonly used in the curing of meats.

FOOD IRRADIATION

Food irradiation involves the use of either high-speed electron beams or high-energy radiation with wavelengths smaller than 200 nanometres, or 2000 angstroms (*e.g.*, X rays and gamma rays). These rays contain sufficient energy to break chemical bonds and ionize molecules that lie in their path. The two most common sources of high-energy radiation used in the food industry are cobalt-60 (^{60}Co) and cesium-137 (^{137}Cs). For the same level of energy, gamma rays have a greater penetrating power into foods than high-speed electrons.

The unit of absorbed dose of radiation by a material is denoted as the gray (Gy), one gray being equal to the absorption of one joule of energy by one kilogram of food. The energy possessed by an electron is called an electron volt (eV). One eV is the amount of kinetic energy gained by an electron as it accelerates through an electric potential difference of one volt. It is usually more convenient to use a larger unit such as megaelectron volt (MeV), which is equal to one million electron volts.

Biological effects of irradiation. Irradiation has both direct and indirect effects on biological materials. The direct effects are due to the collision of radiation with atoms, resulting in an ejection of electrons from the atoms. The indirect effects are due to the formation of free radicals (unstable molecules carrying an extra electron) during the radiolysis (radiation-induced splitting) of water molecules. The radiolysis of water molecules produces hydroxyl radicals, highly reactive species that interact with the organic molecules present in foods. The products of these interactions cause many of the characteristics associated with the spoilage of food, such as off-flavours and off-odours.

Killing of bacteria

Positive effects. The bactericidal (bacteria-killing) effect of ionizing radiation is due to damage of the biomolecules of bacterial cells. The free radicals produced during irradiation may destroy or change the structure of cellular membranes. In addition, radiation causes irreversible changes to the nucleic acid molecules (*i.e.*, DNA and RNA) of bacterial cells, inhibiting their ability to grow. Pathogenic bacteria that are unable to produce resistant endospores in foods such as poultry, meats, and seafood can be eliminated by radiation doses of 3 to 10 kilograys. If the dose of radiation is too low, then the damaged DNA can be repaired by specialized enzymes. If oxygen is present during irradiation, the bacteria are more readily damaged. Doses in the range of 0.2 to 0.36 kilograys are required to stop the reproduction of *Trichinella spiralis* (the parasitic worm that causes trichinosis) in pork, although much higher doses are necessary to eliminate it from the meat.

The dose of radiation used on food products is divided into three levels. Radappertization is a dose in the range of 20 to 30 kilograys, necessary to sterilize a food product. Radurization is a dose of 1 to 10 kilograys, that, like pasteurization, is useful for targeting specific pathogens. Radicidation involves doses of less than 1 kilogray for extending shelf life and inhibiting sprouting.

Negative effects. In the absence of oxygen, radiolysis of lipids leads to cleavage of the interatomic bonds in the fat molecules, producing compounds such as carbon dioxide, alkanes, alkenes, and aldehydes. In addition, lipids are highly vulnerable to oxidation by free radicals, a process that yields peroxides, carbonyl compounds, alcohols, and lactones. The consequent rancidity, resulting from the irradiation of high-fat foods, is highly destructive to their sensory quality. To minimize such harmful effects, fatty foods must be vacuum-packaged and held at subfreezing temperatures during irradiation.

Proteins are not significantly degraded at the low doses of radiation employed in the food industry. For this reason irradiation does not inactivate enzymes involved in food spoilage, as most enzymes survive doses of up to 10 kilograys. On the other hand, the large carbohydrate molecules that provide structure to foods are depolymerized (broken down) by irradiation. This depolymerization reduces the gelling power of the long chains of structural carbohydrates. However, in most foods some protection against these deleterious effects is provided by other food constituents. Vitamins A, E, and B$_1$ (thiamine) are also sensitive to irradiation. The nutritional losses of a food product are high if air is not excluded during irradiation.

Safety concerns. Based on the beneficial effects of irradiation on certain foods, several countries have permitted its use for specific purposes, such as the inhibition of sprouting of potatoes, onions, and garlic; the extension of shelf life of strawberries, mangoes, pears, grapes, cherries, red currants, and cod and haddock fillets; and the insect disinfestation of pulses, peanuts, dried fruits, papayas, wheat, and ground-wheat products.

Large-scale studies conducted around the world have concluded that irradiation does not cause harmful reactions in foods. In 1980 a joint committee of the Food and Agriculture Organization (FAO), the International Atomic Energy Agency (IAEA), and the World Health Organization (WHO) declared that an overall average dose of radiation of 10 kilograys was safe for food products. The maximum energy emitted by ^{60}Co and ^{137}Cs is too low to induce radioactivity in food. The energy output of electron-beam generators is carefully regulated, and the recommended energy outputs are too low to cause radioactivity in foods.

PACKAGING

Because packaging helps to control the immediate environment of a food product, it is useful in creating conditions that extend the storage life of a food. Packaging materials commonly used for foods may be classified as flexible (paper, thin laminates, and plastic film), semirigid (aluminum foil, laminates, paperboard, and thermoformed plastic), and rigid (metal, glass, and thick plastic). Plastic materials are widely used in food packaging because they are relatively cheap, lightweight, and easy to form into desired shapes.

Table 2: Nutrient Composition of Selected Raw Cereal Grains
(per 100 grams)

Cereal grain	energy (kcal)	water (g)	carbohydrate (g)	protein (g)	fat (g)	minerals (g)
Barley (pearled)	352	10.09	77.72	9.91	1.16	1.11
Corn (field)	365	10.37	74.26	9.42	4.74	1.20
Millet	378	8.67	72.85	11.02	4.22	3.25
Oats (oatmeal)	384	8.80	67.00	16.00	6.30	1.90
Rice (brown; long-grain)	370	10.37	77.24	7.94	2.92	1.53
Rye	335	10.95	69.76	14.76	2.50	2.02
Sorghum	339	9.20	74.63	11.30	3.30	1.57
Wheat (hard red winter)	327	13.10	71.18	12.61	1.54	1.57

Source: *Composition of Foods,* Agriculture Handbook no. 8–20, U.S. Department of Agriculture.

Barrier properties of packaging materials

The selective permeability of polymer-based materials to gases, such as carbon dioxide and oxygen, as well as light and moisture, has led to the development of modified-atmosphere packaging. If the barrier properties are carefully selected, a packaging material can maintain a modified atmosphere inside the package and thus extend the shelf life of the food product.

Dehydrated foods must be protected from moisture during storage. Packaging materials such as polyvinyl chloride, polyvinylidene chloride, and polypropylene offer low moisture permeability. Similarly, packaging materials with low gas permeability are used for fatty foods in order to minimize oxidation reactions. Because fresh fruits and vegetables respire, they require packaging materials, such as polyethylene, that have high permeability to gases.

Smart packages offer properties that meet the special needs of certain foods. For example, packages made with oxygen-absorbing materials remove oxygen from the inside of the package, thus protecting oxygen-sensitive products from oxidation. Temperature-sensitive films exhibit an abrupt change in gas permeability when they are subjected to a temperature above or below a set constant. These films change from a crystalline structure to an amorphous structure at a set temperature, causing the gas permeability to change substantially.

STORAGE

Food storage is an important component of food preservation. Many reactions that may deteriorate the quality of a food product occur during storage. The nutrient content of foods may be adversely affected by improper storage. For example, a significant amount of vitamin C and thiamine may be lost from foods during storage. Other undesirable quality changes that may occur during storage include changes in colour, development of off-flavours, and loss of texture. A properly designed food storage system allows fresh or processed foods to be stored for extended duration while maintaining quality.

The most important storage parameter is temperature. Most foods benefit from storage at a constant, low temperature where the rates of most reactions decrease and quality losses are minimized. In addition, foods containing high concentrations of water must be stored in high-humidity environments in order to prevent the excessive loss of moisture.

Careful control of atmospheric gases, such as oxygen, carbon dioxide, and ethylene, is important in extending the storage life of many products. For example, in the United States and Canada the apple industry utilizes controlled-atmosphere storage facilities in order to preserve the quality of the fruit. Use of controlled atmospheres to increase the shelf life of fruits was first shown in 1819 by Jacques-Étienne Berard, a professor at the School of Pharmacy at Montpellier, Fr. The commercial development of this technique occurred more than 100 years later with the pioneering work of Franklin Kidd and Cyril West at the Low Temperature Research Station at Cambridge, Eng.

(N.W.D./R.P.Si.)

Cereals and other starch products

Cereals, or grains, are members of the grass family cultivated primarily for their starchy seeds (technically, dry fruits), which are used for human food, for livestock feed, and as a source of industrial starch. Wheat, rice, corn (maize), rye, oats, barley, sorghum, and some of the millets are common cereals; their composition is shown in Table 2.

Starch, a carbohydrate stored in most plants, is a major constituent of the average human diet, providing a low-cost energy source with good keeping qualities. Cereals are high in starch, which may be used in pure or flour form. Starches are also obtained from such root sources as potatoes and from the pith of tropical palm trees. Various starches are used commercially in food processing and in the manufacture of laundering preparations, paper, textiles, adhesives, explosives, and cosmetics.

This section treats the processing and utilization of the major cereals—wheat, rice, barley, rye, oats, corn, sorghum, millet, and buckwheat; of important starchy foods consumed in certain countries instead of cereals, including potatoes and cassava; and of soybeans, legumes widely used in the bakery industry. Wheat species are treated in detail, other cereals in a more general way.

CEREAL PROCESSING AND UTILIZATION

Milling. Cereal processing is complex. The principal procedure is milling—that is, the grinding of the grain so that it can be easily cooked and rendered into an attractive foodstuff. Cereals usually are not eaten raw, but different kinds of milling (dry and wet) are employed, depending on the cereal itself and on the eating customs of the consumer. Wheat may be crushed with grinding stones or similar devices or by modern automated systems employing steel cylinders, followed by air purification and numerous sievings to separate the endosperm from the outer coverings and the germ.

Corn is often milled by wet processes, but dry milling is also practiced, especially in the developing countries. Corn, with its high germ content, is inclined to respire more during storage and, unless precautions are taken, may increase in temperature during incorrect storage. Most other cereals are ground in the dry state. Some cereal grains are polished, removing most of the bran and germ and leaving the endosperm.

Uses. *Human food.* Cereals are used for both human and animal food and as an industrial raw material. Although milled white flour is largely used for bread production, especially in industrialized countries, the grain may be converted to food in other ways. In India the major part of the grain is not ground into flour in roller mills but is roughly ground in small crushing mills into a meal called *atta.* This meal is cooked into flat cakes known as chapatis.

Animal food. The principal cereals used as components of animal feeds are wheat and such wheat by-products as the outer coverings separated in the preparation of white flour (bran and the more floury middlings), corn, barley, sorghum, rye, and oats. These are supplemented by protein foods and green fodders.

Animal foods require proper balance between the cereals (carbohydrates) and the more proteinous foods, and they must also contain suitable amounts of necessary minerals, vitamins, and other nutrients. The compounded ration for a milking cow generally contains about 50–80 percent cereals, consisting of wheat by-products, flaked or ground corn, barley, sorghum, wheat, and oats. Requirements for most balanced rations for pigs and poultry are similar. Corn is especially useful in high-energy feeds either as meal or as the flaked and partly gelatinized product; barley is desirable for fattening, and oats help provide a better balanced cereal for livestock. Without cereals for use in farm animal foods, the available supply of the animal protein required in the human diet would be greatly reduced.

Industrial uses. The relatively minor use of cereals in nonfood products includes the cellulose in the straw of cereals by the paper industry, flour for manufacturing sticking pastes and industrial alcohol, and wheat gluten for core binders in the casting of metal. Rice chaff is often used as fuel in Asia.

WHEAT: VARIETIES AND CHARACTERISTICS

The three principal types of wheat used in modern food production are *Triticum vulgare* (or *aestivum*), *T. durum*, and *T. compactum*. *T. vulgare* provides the bulk of the wheat used to produce flour for bread making and for cakes and biscuits (cookies). It can be grown under a wide range of climatic conditions and soils. Although the yield varies with climate and other factors, it is cultivated from the southernmost regions of America almost to the Arctic and at elevations from sea level to over 10,000 feet. *T. durum*, longer and narrower in shape than *T. vulgare*, is mainly ground into semolina (purified middlings) instead of flour. Durum semolina is generally the best type for the production of pasta foods. *T. compactum* is more suitable for confectionery and biscuits than for other purposes.

Wheat grain composition

The wheat grain, the raw material of flour production and the seed planted to produce new plants, consists of three major portions: (1) the embryo or germ (including its sheaf, the scutellum) that produces the new plant, (2) the starchy endosperm, which serves as food for the germinating seed and forms the raw material of flour manufacture, and (3) various covering layers protecting the grain. (See Figure 3.) Although proportions vary, other cereal grains follow the same general pattern. Average wheat grain composition is approximately 85 percent endosperm, 13 percent husk, and 2 percent embryo.

Characteristic variations of the different types of wheat are important agricultural considerations. Hard wheats include the strong wheats of Canada (Manitoba) and the similar hard red spring (HRS) wheats of the United States. They yield excellent bread-making flour because of their high quantity of protein (approximately 12–15 percent), mainly in the form of gluten. Soft wheats, the major wheats grown in the United Kingdom, most of Europe, and Australia, result in flour producing less attractive bread than that achieved from strong wheats. The loaves are generally smaller, and the crumb has a less pleasing structure. Soft wheats, however, possess excellent characteristics for the production of flour used in cake and biscuit manufacture.

Wheats intermediate in character include the hard red winter (HRW) wheats of the central United States and wheat from Argentina. There are important differences between spring and winter varieties. Spring wheats, planted in the early spring, grow quickly and are normally harvested in late summer or early autumn. Winter wheats are planted in the autumn and harvested in late spring or early summer. Both spring and winter wheats are grown in different regions of the United States and Russia. Winter varieties can be grown only where the winters are sufficiently mild. Where winters are severe, as in Canada, spring types are usually cultivated, and the preferred varieties mature early, allowing harvesting before frost.

Strong and weak flour

In baking and confectionery, the terms strong and weak indicate flour from hard and soft wheats, respectively. The term strength is used to describe the type of flour, strong flours being preferred for bread manufacture and weak flours for cakes and biscuits. Strong flours are high in protein content, and their gluten has a pleasing elasticity; weak flours are low in protein, and their weak, flowy gluten produces a soft, flowy dough. (See Figure 5.)

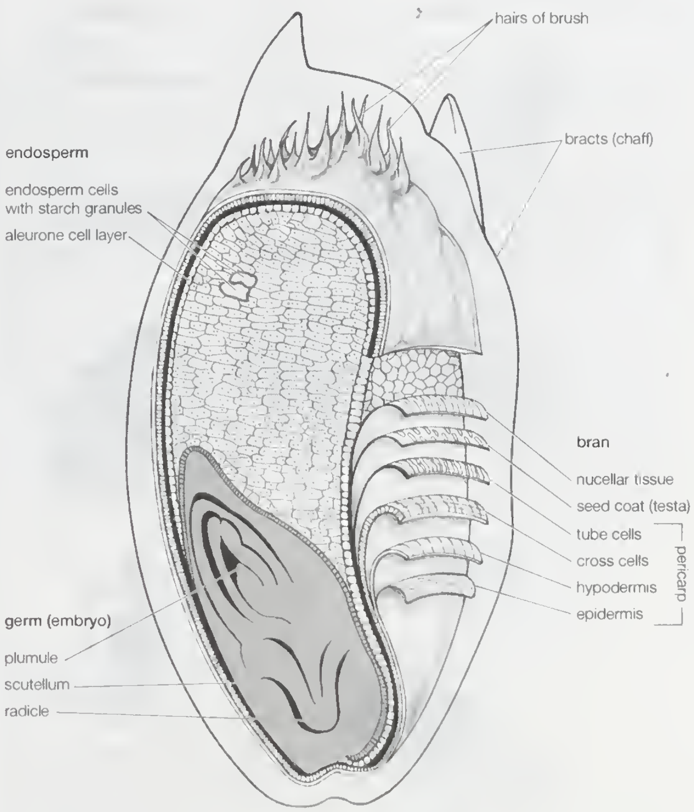


Figure 3: The outer layers and internal structures of a kernel of wheat.
Encyclopædia Britannica Inc

Wheat breeders regularly produce new varieties, not only to combat disease but also to satisfy changing market demands. Many varieties of wheat do not retain their popularity, and often those popular in one decade are replaced in the next. New varieties of barley have also been developed, but there have been few varieties of rice.

WHEAT FLOUR

The milling of wheat into flour for the production of bread, cakes, biscuits, and other edible products is a huge industry. Cereal grains are complex, consisting of many distinctive parts. The objective of milling is separation of the floury edible endosperm from the various branny outer coverings and elimination of the germ, or embryo. Because wheats vary in chemical composition, flour composition also varies.

Although some important changes have occurred in flour milling, basic milling procedure during the past 100 years has employed the gradual reduction process as described below and illustrated in Figure 4.

Milling. In modern milling considerable attention is given to preliminary screening and cleaning of the wheat or blend of wheats to exclude foreign seed and other impurities. The wheat is dampened and washed if it is too dry for subsequent efficient grinding, or if it is too damp it is gently dried to avoid damaging the physical state of the protein present, mainly in the form of the elastic substance gluten.

Gradual reduction from grain to flour

The first step in grinding for the gradual reduction process is performed between steel cylinders, with grooved surfaces, working at differential speeds. The wheat is directed between the first "break," or set of rolls, and is partially torn open. There is little actual grinding at this stage. The "chop," the resulting product leaving the rolls, is sieved, and three main separations are made: some of the endosperm, reduced to flour called "first break flour"; a fair amount of the coarse nodules of floury substances from the endosperm, called semolina; and relatively large pieces of the grain with much of the endosperm still adhering to the branny outsides. These largish portions of the wheat are fed to the second break roll. The broad

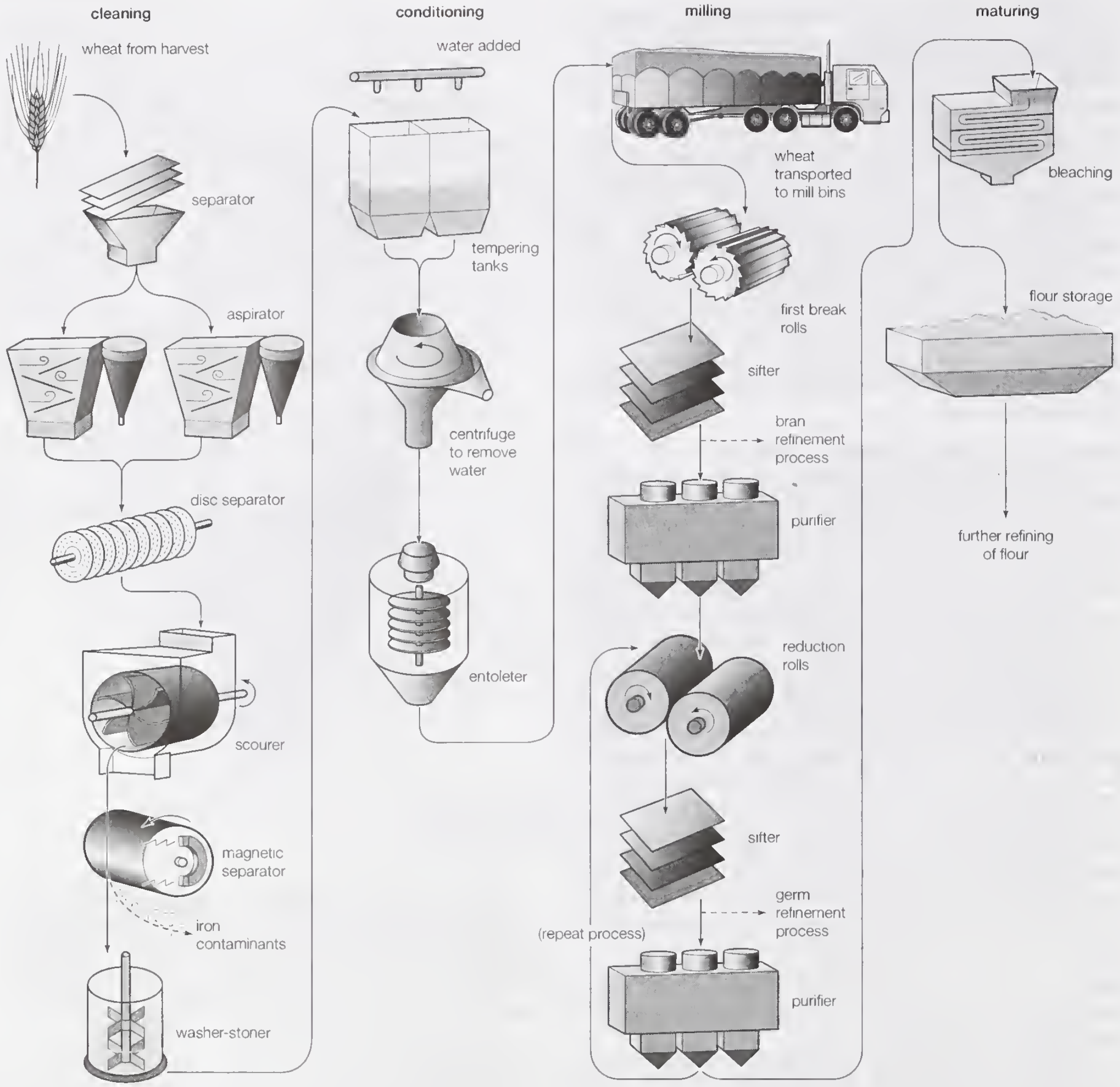


Figure 4: Stages in the milling of wheat into flour.
Encyclopædia Britannica, Inc.

objective of this gradual reduction process is the release, by means of the various sets of break rolls, of inner endosperm of the grain, in the form of semolina, in amounts sufficient that the various semolinas from four or five break rolls can be separated by suitable sieving and the branny impurities can be removed by air purifiers and other devices. The cleaned semolinas are reduced to fine flour by grinding between smooth steel rolls, called reduction rolls. The flour produced in the reduction rolls is then sieved out. There are usually four or five more reduction rolls and some "scratch" rolls to scrape the last particles of flour from branny stocks. Since the various sieving and purification processes free more and more endosperm in the form of flour, flour is obtained from a whole series of processing operations. The flour is sieved out after each reduction roll, but no attempt is made to reduce to flour all the semolina going to a particular reduction roll. Some of the endosperm remains in the form of finer semolina and is again fed to another reduction roll. Each reduction roll tends to reduce more of the semolina to flour and to flatten bran particles and thus facilitate the sieving out of the branny fractions. The sieving plant generally employs machines called plan-sifters, and the air purifiers also produce a whole series of floury stocks.

Modern flour processing consists of a complicated series of rolls, sieves, and purifiers. Approximately 72 percent of the grain finally enters the flour sack.

The sacked flour may consist of 20 or more streams of flour of various states of purity and freedom from branny specks. By selection of the various flour streams it is possible to make flour of various grades. Improvements in milling techniques, use of newer types of grinding machinery in the milling system, speeding up of rolls, and improved skills have all resulted in flour produced by employing the fundamentals of the gradual reduction process but with simplified and shorter milling systems. Much less roll surface is now required than was needed as recently as the 1940s.

Patent flour

The purest flour, selected from the purest flour streams released in the mill, is often called patent flour. It has very low mineral (or ash) content and is remarkably free from traces of branny specks and other impurities. The bulk of the approximately 72 percent released is suited to most bread-making purposes, but special varieties are needed for some confectionery purposes. These varieties may have to be especially fine for production of specialized cakes, called high-ratio cakes, that are especially light and have good keeping qualities.

In many countries the flour for bread production is submitted to chemical treatments to improve the baking quality.

In modern processing, regrinding of the flour and subsequent separation into divisions by air treatment has enabled the processors to manufacture flour of varying protein content from any one wheat or grist of wheats.

Composition and grade. Flour consists of moisture, proteins (mainly in gluten form), a small proportion of fat or lipids, carbohydrates (mainly starch, with a small amount of sugar), a trace of fibre, mineral matter (higher amounts in whole meal), and various vitamins. Composition varies among the types of flour, semolinas, middlings, and bran (see Figure 5).

Gluten

Protein content. For bread making it is usually advantageous to have the highest protein content possible (depending on the nature of the wheat used), but for most other baked products, such as cookies (sweet biscuits) and cakes, high protein content is rarely required. Gluten can easily be washed out of flour by allowing a dough made of the flour and water to stand in water a short time, followed by careful washing of the dough in a gentle stream of water, removing the starch and leaving the gluten. For good bread-making characteristics, the gluten should be semi-elastic, not too stiff and unyielding but not soft and flowy, although a flowy quality is required for biscuit manufacture.

The gluten, always containing a small amount of adhering starch, is essentially hydrated protein. With careful drying it will retain its elasticity when again mixed with water and can be used to increase the protein content of specialized high-protein breads.

Sometimes locally grown wheat, often low in protein, may be the only type available for flour for bread making. This situation exists in parts of France, Australia, and South Africa. The use of modern procedures and adjustment of baking techniques, however, allow production of satisfactory bread. In the United Kingdom, millers prefer a blend of wheat, much of it imported, but modern baking procedures have allowed incorporation of a larger proportion of the weak English wheat than was previously feasible.

Treatment of flour. Use of "improvers," or oxidizing substances, enhances the baking quality of flour, allowing production of better and larger loaves. Relatively small amounts are required, generally a few parts per million. Although such improvers and the bleaching agents used to rectify excessive yellowness in flour are permitted in most countries, the processes are not universal. Improvers include bromates, chlorine dioxide (in gaseous form), and azodicarbonamide. The most popular bleacher used is benzoyl peroxide.

Grade. The grade of flour is based on freedom from branny particles. Chemical testing methods are employed to check general quality and particularly grade and purity. Since the ash (mineral content) of the pure branny coverings of the wheat grain is much greater than that of the pure endosperm, considerable emphasis is placed on use of the ash test to determine grade. Bakers will generally pay higher prices for pure flour of low ash content, as the flour is brighter and lighter in colour. Darker flours may have ash content of 0.7 to 0.8 percent or higher.

Encyclopædia Britannica, Inc

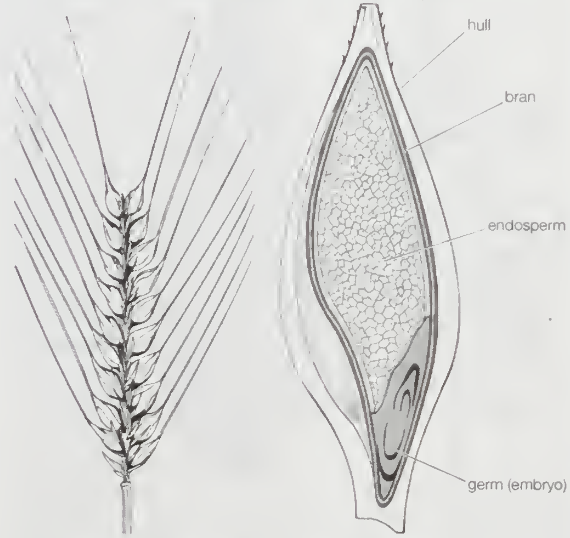


Figure 6: (Left) The barley spike, with rows of barley florets. (Right) Cross section of the barleycorn.

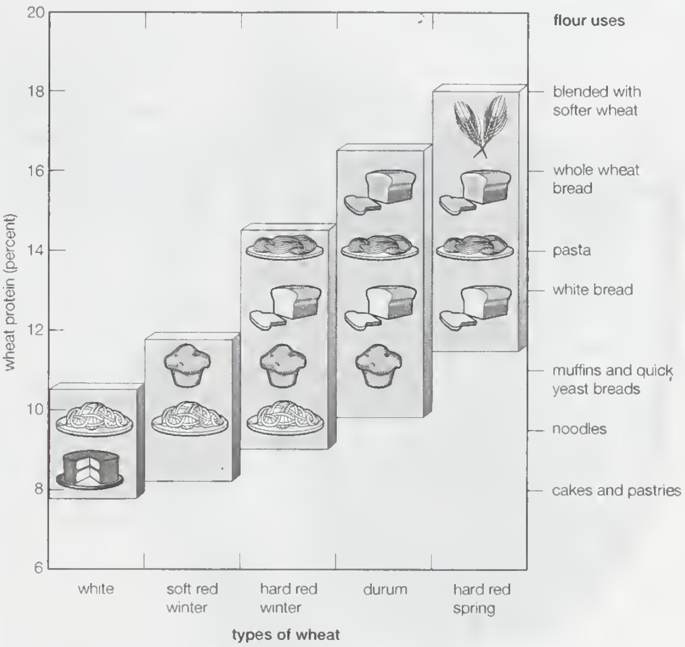


Figure 5: The protein content and major food uses of certain varieties of wheat.

Source: W.G. Heid, *U.S. Wheat Industry*, Agricultural Economics Report, No. 432, U.S. Department of Agricultural Economics, Statistics and Cooperative Service

A widely employed modern method for testing flour colour is based on the reflectance of light from the flour in paste form. This method requires less than a minute; the indirect ash test requires approximately one to two hours.

NONWHEAT CEREALS

Barley. Most of the barley grown in the world is used for animal feed, but a special pure barley is the source of malt for beer production. Barley is also used in the manufacture of vinegar, malt extract, some milk-type beverages, and certain breakfast foods. In addition, in flaked form it is employed in some sections of the brewing industry, and pearl barley (skins removed by emery friction) is used in various cooked foods.

Barley (see Figure 6) can be cultivated on poorer soil and at lower temperatures than wheat. An important characteristic in barley is "winterhardiness," which involves the ability to modify or withstand many types of stresses, particularly that of frost. However, barley is subject to many of the diseases and pests that affect wheat.

The use of barley in animal feed is increasing; it has been a basic ingredient of pig foods for years and is increasingly used for cattle feed. Its use in poultry foods has decreased because it has a lower starch equivalent when compared with wheat or corn and thus provides a lower-energy ration, unsuitable in modern poultry production. Barley vitamin content is similar to that of wheat.

Flint and dent corn

Corn. Corn, or maize, a cereal cultivated in most warm areas of the world, has many varieties. The United States, the principal producer of corn, cultivates two main commercial types, *Zea indurata* (flint corn) and *Z. indentata* (dent corn). The plant grows to a height of about three metres or more. The corn kernel (shown in Figure 7) is large for a cereal, with a high embryo content, and corn oil extracted from the germ is commercially valuable. The microscopic appearance of the starch is distinctive, and the principal protein in ordinary corn is the prolamin zein, constituting half of the total protein. On hydrolysis zein yields only very small amounts of tryptophan or lysine, making it low in biological value. The proteins of corn, like those of most cereals other than wheat, do not provide an elastic gluten.

Much of the corn is wet-processed to produce corn flour, widely used in cooking (see below *Starch products: Cornstarch*). Corn, dry-milled as grits or as meal or turned into flaked corn with some of its starch partially gelatinized, is a popular component in compounded animal feedstuffs. In dry-milled form it is also the basis of human food throughout large areas of Africa and South America. Its

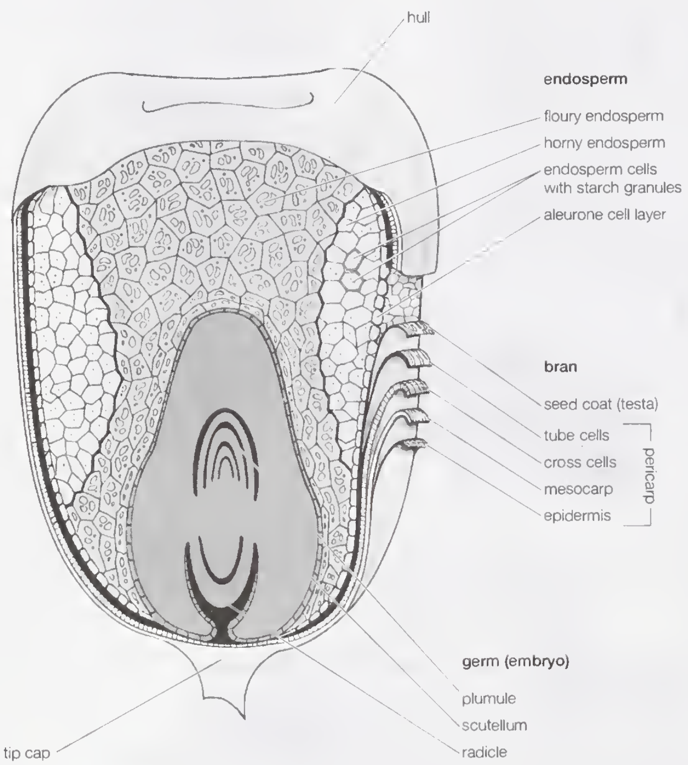


Figure 7: The outer layers and internal structures of a kernel of corn.
Encyclopædia Britannica, Inc.

nutritive value is limited by its low lysine content. Much recent research has involved development of a corn with higher lysine content. Mutants have been produced containing much less zein but possessing protein with higher than normal lysine and tryptophan contents, sometimes increased as high as 50 percent. These corns, called Opaque-2 and Floury-2, possess certain drawbacks. They are generally lower in yield than dent hybrids, are subject to more kernel damage when combine-harvested, and may be more difficult to process. Nevertheless, these new hybrid corns are expected to become widely cultivated, and the principles involved in their production may also be applied to sorghum, wheat, and rice. Corn is popular for use in breakfast foods.

Sorghum. Sorghum, also called milo, is of smaller size than corn but is generally the same type of cereal, with similar appearance. Its numerous types are mainly used for animal feeding. It is grown extensively in the United States, Pakistan, central India, Africa, and China. In the sorghum endosperm, the proteins soluble in hot 60 percent alcohol, called kafirin, constitute the major portion of the protein. Milo germ oil is similar to corn germ oil; its major fatty acids are palmitic, stearic, and particularly oleic and linoleic. Milo is commercially graded in the United States. In waxy varieties the starch is principally in the form of amylopectin, with very little amylose. Such starches possess special viscosity characteristics.

Oats. Oats belong to the botanical genus *Avena,* which includes a large number of types, the principal being *A. sativa, A. sterilis,* and *A. strigosa.* Oats are widely grown in most countries but are not suitable for Mediterranean climates. Oats are frequently grown on farms as feed for the farm's livestock. They are well balanced chemically, with fairly high fat content, and are particularly suitable for feeding horses and sheep.

Although a large portion of the world's oat production is used for animal feed, oatmeal is a popular human food in many countries. Thin-skinned grains, fairly rich in protein and not too starchy, are selected (see Figure 8). Preliminary cleaning is essential for human consumption. The oats are then kilned (roasted). Thin-husked oats yield 60 percent oatmeal; varieties with thick husks yield only 50 percent.

Rapid development of rancidity is a serious problem in oats and oat products. The free fatty acid content must be controlled because formation of these acids tends to produce a soapy taste resulting from the activity of the enzyme lipase. A few minutes of steam treatment normally destroys the lipase activity in the grain.

Rye. Rye, which has been known for some 2,000 years, ranks second to wheat as a bread flour. The principal rye producers are Russia, Poland, Belarus, Germany, and Ukraine. The popularity of true rye bread is decreasing, and a similar bread, retaining some of the original characteristics, is now made from a rye and wheat blend. The protein of European rye tends to be low and does not yield gluten in the same way as does wheat. Rye bread, closer-grained and heavier than wheat bread, is aerated by the use of a leaven (sourdough) rather than yeast. The grain is susceptible to attack by the parasitic fungus ergot (*Claviceps purpurea*).

Rice. Cultivated rice is known botanically as *Oryza sativa,* only one of some 25 species comprising the genus *Oryza*. The importance of this cereal to certain parts of the world may be seen from the fact that in Sanskrit there exists, besides the usual word for rice, another term signifying "sustainer of the human race." Rice is the staple food for millions in Southeast Asia, almost equal to wheat in importance among the world's cereal crops.

Rice as a staple food

Cultivation. More than 90 percent of the world's rice is grown in Asia, principally in China, India, Indonesia, and Bangladesh, with smaller amounts grown in Japan, Pakistan, and various Southeast Asian nations. Rice is also cultivated in parts of Europe, in North and South America, and in Australia. The bulk of the rice cultivated in Asia is grown under water in flooded fields. Successful production depends on adequate irrigation, including construction of dams and waterwheels, and on the quality of the soil. Long periods of sunshine are essential. Rice yields vary considerably, ranging from 700 to 4,000 kilograms per hectare (600 to 3,500 pounds per acre). Adequate irrigation, which means inundation of the fields to a depth of several inches during the greater part of the growing season, is a basic requirement for productive land use.

Dryland paddy production, with harvesting by modern mechanical means, is limited to a few areas, and it produces only a fraction of the total world crop.

As with other cereals, weeds, especially wild red rice, are a constant problem. The commonest pests include plant bugs, stem borers, worms, and grasshoppers. The crop, often harvested with a sickle, is frequently dried in earth or concrete pits. Threshing is often carried out by trampling or with crude implements. Only in a few rice-growing

Encyclopædia Britannica, Inc.

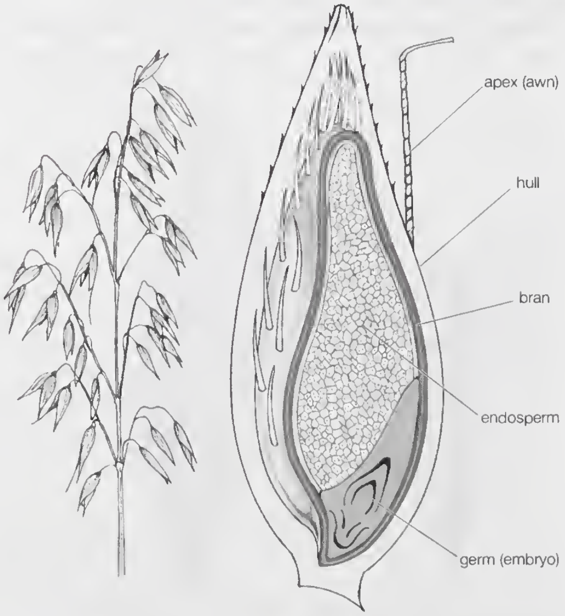


Figure 8: (Left) The oat panicle, bearing multiple oat florets. (Right) Cross section of the oat grain.

regions are more modern procedures used in harvesting.

Manpower requirements for crops vary enormously, but over 400 man-hours per acre are required in smallholdings in Asia, where labour is cheap.

In Asia the paddy is cultivated in three main types of soil, including clays with a firm bottom within a few inches of the surface; silts and soft clays with soft bottoms becoming hard on drying; and peats and "mucks" containing peat, provided the depth of the peat is not excessive. Fields must be drained and dried before harvesting. When combine harvesters or binder threshers are employed, the grain must be dried to about 14 percent moisture so that no deterioration takes place in storage. When reaper binders are used, the crop is "shocked" in certain ways so that the grain is protected from rain.

Milling. Milling methods used in most of Asia are primitive, but large mills operate in Japan and some other areas. Hulling of the paddy is usually accomplished by pestle and mortar worked by hand, foot, or water power. Improvements are slowly taking place. The yield of milled rice is dependent on the size and shape of the grain, the degree of ripeness, and the extent of exposure to the sun. Some large mills, handling 500 to 1,000 tons of paddy daily, have specialized hulling plants with consequent smaller losses from broken grain. They generally employ modern milling techniques and rely on controlled drying plants instead of on sun drying.

Yield of the rice grain

The weight of the husk is about 20 percent of the weight of the paddy, and there are losses of about 5 percent from dirt, dead grains, and other impurities. Approximately 74 percent of the paddy is available as rice and rice byproducts. The yield from milling and subsequent emery polishings includes about 50 percent whole rice, 17 percent broken rice, 10 percent bran, and 3 percent meal. Rice grains have a series of thin coats that can be removed or partially removed in the process of pearling and whitening (see Figure 9).

About 60 percent of the Indian rice is parboiled. In the parboiling process the paddy is steeped in hot water, subjected to low-pressure steam heating, then dried and milled as usual. Parboiling makes more rice available from the paddy, and more nutrients (largely vitamin B_1) are transferred from the outer coverings to the endosperm, improving the nutritive value of the finished product. Parboiled rice may contain two to four times as much thiamine (vitamin B_1) and niacin as milled raw rice, and losses in cooking may also be reduced.

Alcoholic drinks, such as sake in Japan and *wang-tsin* in China, are made from rice with the aid of fungi. The hull or husk of paddy, of little value as animal feed because of a high silicon content that is harmful to digestive and respiratory organs, is used mainly as fuel.

Nutritive value. The lysine content of rice is low. As rice is not a complete food, and the majority of Asians live largely on rice, it is important that loss of nutrients in processing and cooking should be minimal. Lightly milled rice has about 0.7 milligram of vitamin B_1 per 1,000 nonfatty calories, and the more costly highly milled product has only 0.18 milligram of B_1 on the same basis. For adequate nutrition, vitamin B_1 in the daily diet on this basis should be 0.5–0.6 milligram. The amount of fat-soluble vitamins in rice is negligible.

In some countries rice is enriched by addition of synthetic vitamins. According to U.S. standards for enriched rice, each pound must contain 2–4 milligrams of thiamine, 1.2–2.4 milligrams of riboflavin, 16–32 milligrams of niacin, and 13–26 milligrams of iron. In enriched rice the loss of water-soluble vitamins in cooking is much reduced because enrichment is applied to about 1 grain in 200, and these enriched grains are protected by a collodion covering. In ordinary rice, especially when open cookers are employed or excessive water is used, nutrient losses can be high.

Millet. This term is applied to a variety of small seeds originally cultivated by the ancient Egyptians, Greeks, and Romans and still part of the human diet in China, Japan, and India, though in Western countries it is used mainly for birdseed. The genus is termed *Panicum.* The small seed is normally about two millimetres long and nearly two millimetres broad. The term *proso* is one of several alternative names. Japanese barnyard millet is a well-known variety.

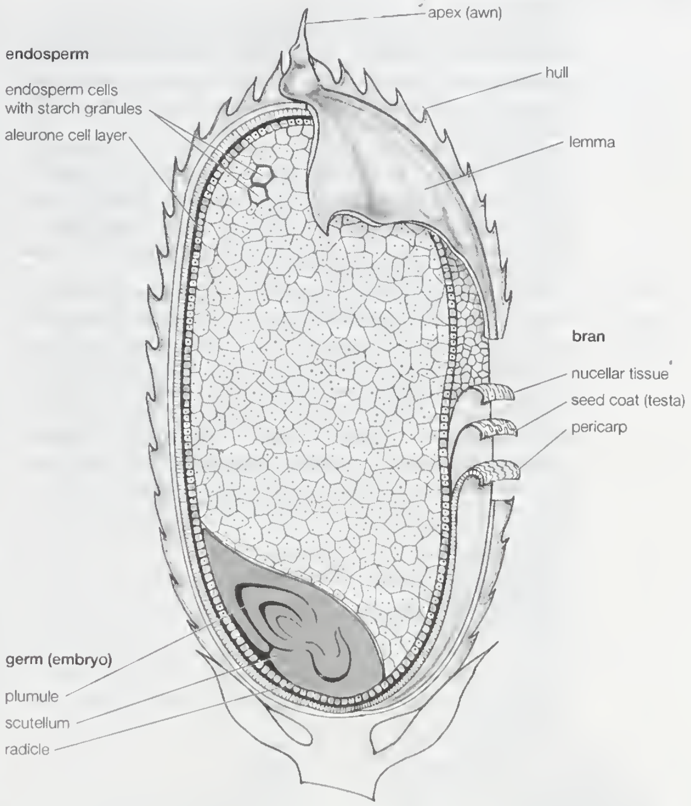


Figure 9: The outer layers and internal structures of a rice grain.

Encyclopædia Britannica, Inc

OTHER STARCH-YIELDING PLANTS

Cassava. Cassava, often called manioc, is not a cereal but a tuber; however, it replaces cereals in certain countries, supplying the carbohydrate content of the diet. The botanical name is *Manihot esculenta,* and the plant is native to South America, especially Brazil. It is now grown in Indonesia, Malaysia, the Philippines, Thailand, and parts of Africa. A valuable source of starch, cassava is familiar in many developed countries in a granular form known as tapioca.

Tapioca

Easily cultivated and curiously immune to most food-crop pests, cassava is a staple crop in several areas of Latin America. The actual tubers may weigh up to 14 kilograms (30 pounds). Some tubers may be bitter and contain dangerously large amounts of prussic acid.

Dry milling of cassava is rarely practiced because it yields a product inferior to wet-processed starch in which the tubers are crushed or rasped with water and the starch is permitted to settle. Wet starch is dried to a point where it can be crumbled by pressing it through metal plates (or sieves). This crumbled material is subjected to a rotary motion, sometimes carried out on canvas cloth fastened to cradle-shaped frames. Another method is to tumble the material in revolving steam-jacketed cylinders so that the particles assume a round pellet form and are partially gelatinized as they dry. Sun drying is employed in both homes and small mills.

Many tapioca factories and mills are equipped with modern raspers, special shaking or rotating sieves, and settling tanks of various types; but some fermentation takes place, and small rural mills can often be identified by the smell of butyric acid. In larger mills, centrifuges are replacing the settling tanks.

For its various industrial uses, the tuber usually goes under its alternative name, manioc. It is used in the textile industries, explosives manufacture, leather tanning, and production of glues and dextrins and alcohol.

Fresh cassava leaves are rich in protein, calcium, and vitamins A and C. Their prussic acid level must be reduced to safe limits by boiling; the duration of boiling depends on the variety of the leaves. Cassava leaves are a popular vegetable in Africa, and the tuber also is used in meal for animal feed.

Soybean. Soybean (*Glycine max*) is not a cereal but a legume; because of its widespread use in the baking industry, it may appropriately be dealt with here. Soybean provides protein of high biological value. Although Asia is its original source, the United States became the major world producer in the late 20th century.

The valuable oil of the soybean, widely used in industry, is extracted either by solvents or by expellers. The amino acid distribution of soy protein is more like that found in animal protein than the protein from most vegetable sources; for example, lysine comprises about 5.4 percent. The oil content includes useful amounts of phosphorus; the phosphatide content of soy flour is about 2 percent and is a mixture of lecithin and cephalin. The low carbohydrate content exists mainly as sugars. Table 3 shows the amino acid composition of soy protein.

Table 3: Amino Acid Composition of Soy Protein
(calculated to 16 percent nitrogen)

	percent
Arginine	5.8
Histidine	2.3
Lysine	5.4
Tyrosine	4.1
Tryptophan	1.2
Phenylalanine	5.7
Cystine	0.9
Methionine	2.0
Threonine	4.0
Leucine	6.6
Isoleucine	4.7
Valine	4.2
Glutamic acid	21.0
Aspartic acid	8.8

Although soybeans are a good source of thiamine, much of this may be lost in processing. Average vitamin contents of soybean (as micrograms per gram) are as follows: thiamine 12, riboflavin 3.5, nicotinic acid 23, pyridoxine 8, pantothenic acid 15, and biotin 0.7.

The bulk of the soybean produced in the United States is used for animal feed; the Asian crop goes principally for human diet.

Soybean milk

Soybean milk is produced and used in the fresh state in China and as a condensed milk in Japan. In both of these preparations, certain antinutritive factors (antitrypsin and soyin) are largely removed. In the Western world most soy products are treated chemically or by heat to remove these antinutritive factors along with the unpopular beany taste. Such processing affects the enzymatic activity in the milk.

Soybean is milled to produce soy flour. The flour is often used in a proportion of less than 1 percent in bakery operations. It stiffens doughs and helps to maintain crumb softness. Unprocessed soy flour, because of its lipoxidase enzyme system, is employed with high-speed mixing to bleach the flour in a dough.

In addition to their use in bread, soy products are used in confectionery, biscuits, macaroni, infant and invalid foods, ice cream, chocolate, sausages, sauces, lemon curd, mayonnaise, meat and fish pastes, certain diabetic foods, and in such nonfood products as paint, paper, textiles, and plastics.

A recent development is the isolation of the soybean proteins for use as emulsifiers and binders in meat products and substitutes. Enzyme-modified proteins provide useful egg-albumen supplement for whipped products.

Buckwheat. Botanically, buckwheat is not a cereal but the fruit of *Fagopyrum esculentum*. Its name is probably derived from its resemblance to beechnut. Believed to have originated in China, the plant grows to a height of about one metre and thrives best in cool, moist climates, although it does not easily tolerate frost. It can be grown on a wide range of soils, and a crop can be obtained within 10–12 weeks of sowing. The seed is dark brown in colour and often triangular in shape. It contains about 60 percent carbohydrate, 10 percent protein, and 15 percent fibre. A white flour can be obtained from the seeds (buckwheat cakes and pancakes are popular in certain areas), and buckwheat meal is also used in animal feed. The whole seed may be fed to poultry and game birds. There is some medical interest in buckwheat as a source of rutin, possibly effective in treatment of increased capillary fragility associated with hypertension in humans.

STARCH PRODUCTS

Commercial starches. Starch has been used for many centuries. An Egyptian papyrus paper dating from 3500 BC was apparently treated with a starch adhesive. The major starch sources are tubers, such as potatoes and cassava, and cereals. Current starch production is considerable. Among the major producing areas, the European countries use both domestic wheat and potatoes and imported corn as the raw material; the United States uses corn and such similar cereals as sorghum; and in South America the cassava plant is the major raw material.

Separated from tubers and cereals, starch is used for conversion into various sugars, and half of the world's separated starch is processed into glucose. Starch is also processed for use in adhesives manufacture. In the food industry starch is used as a thickener in the preparation of cornstarch puddings, custards, sauces, cream soups, and gravies. Starch from tubers and cereals provides the carbohydrate of the human diet.

Large quantities of starch and its derivatives are used in the paper and textile industries.

Potato farina

Starch from tubers. In Germany, The Netherlands, Poland, and a number of other countries, the extraction of the starch from potatoes (sometimes called farina) is a major industry. Some factories produce over 300 tons daily. Processing involves continuous and automatic cleaning of the potatoes, thorough disintegration in raspers or hammer mills, and separation of the fibres from the pulp by centrifugal (rotary) sieves. The resulting starch "milk" contains starch in suspension and soluble potato solids in solution. The starch is separated and washed free from the solubles, the water is removed by centrifugal action, and the damp starch is dried. The flash type of dryer, using hot air, is widely employed for starches derived from both tubers and cereals. Sulfurous acid is generally introduced into the process to prevent the development of various microorganisms.

Potato flour is also produced in Germany and other countries, slices of cleaned potatoes being dried, ground, and sieved. In Germany a "potato sago" is produced. The starch cake obtained from the potatoes is crumbled to produce reasonably uniform-size particles that are rounded by tumbling or similar operations, heated to gelatinize the outside layers of the starch, and then dried.

Potatoes were employed in baking to make the barm, or leaven, before compressed distiller's yeast was available, and they have also been used to supplement limited supplies of wheat flour. The potatoes are cleaned, boiled until soft but not mushy, and mixed, in a proportion of 2 to 3 percent, in the dough.

Modern, ready-to-use, dried and powdered mashed potatoes are popular consumer products.

Cassava and tapioca starches are sometimes partially gelatinized by vacuum drying. Protein impurities are low in commercial starches of potato, sago, and tapioca but as high as 0.2 percent in wheat starch and higher in corn flour.

Corn flour

Cornstarch. Corn is wet-milled to produce corn flour, or cornstarch, desirable for cooking because it forms a paste that sets with a "short" texture and separates from molds more cleanly than do the gels produced by such starches as potato, tapioca, and arrowroot, which are "long," or elastic. In wet milling, the grains are first dry-cleaned so that other cereals and some of the impurities are removed, then steeped in warm water containing sulfur dioxide. This process softens the grains, and the outer skin and the germ are rendered removable. The corn is coarsely ground in "degerminating mills," and the slurry is further wet-ground and sieved to remove all the germ and complete the separation of the starch.

The germ, rich in oil, is eventually dried, and the oil is expelled by pressure, providing an excellent edible oil for culinary use, often replacing olive oil. Corn oil is used for salad oil, margarine, and shortening and for such nonfood items as soap.

The pure starch, held in suspension, was formerly collected by gravity as it flowed down tables, but in modern practice the starch suspension is thickened by the elimination of water by means of machines, and the starch is finally separated by the use of centrifuges. The starch is readily dried without gelatinization taking place.

There is a regular demand for a good grade of corn flour, or cornstarch. Roller-milled corn is still produced for human consumption in Africa and elsewhere. In the United States some corn grits are used by brewers, but the bulk of the corn grown is used for animal feed as meal, grits, or in partially gelatinized flake form.

Rice starch. Rice starch, largely used in laundry work, is normally prepared from broken white rice. The broken grains are steeped for several hours in a caustic soda solution, and the alkali is finally washed away with water. The softened grains are ground with more caustic soda solution, and the resulting mass is settled or submitted to centrifugation in a drum. The starch layer is agitated with water (often with 0.25 percent formaldehyde solution added), and the resulting starch liquor is dewatered, washed on a continuous rotary vacuum filter, resuspended in water, and finally dewatered in a perforated basket centrifuge to about 35 percent moisture. In modern processing it is usual to roll out a thick layer of moist starch, which is then slowly dried and falls to pieces as crystals.

Starch composition. Starch consists of two components: amylose and amylopectin. The relative proportion of these two components varies, and they react differently to enzymatic attack. The enzyme β-amylase (maltogenic) attacks the straight chain amylose but is unable to attack most of the branch chain amylopectin. If only β-amylase is present, maltose is produced, together with a residue of the amylopectin portion, or dextrin of high molecular weight. When α-amylase (dextrinogenic) attacks starch, gummy dextrins of low molecular weight are formed and can produce a sticky crumb in bread.

In bread making there is only limited time for such enzymatic attacks on the starch, and only the "attackable" or "damaged" granules can produce the fermentable sugar for the dough. The β-amylase has little effect on viscosity. The viscosity of gelatinized starch is markedly reduced by α-amylase, however, and is therefore valuable in syrup and dextrose manufacture.

The gelatinization of starch that occurs in hot water is an important characteristic, and the viscous pastes formed are influenced by the treatment the starch has received in its preliminary separation from the cereal or tuber. Chemicals affect degree and speed of gelatinization and the nature and viscosity of the pastes formed.

In certain cereals, particularly in special corns, the starch consists almost entirely of amylopectin, and the term "waxy" is applied to such cereals. They are useful for their unusual physical properties and viscosities. They possess outstanding paste clarity, high water-binding capacity, and resistance to gel formation and retrogradation; they are helpful in production of salad dressings, sauces, and pie fillings and in some canned goods; they are useful because of resistance to irreversible gel formation and syneresis on freezing and especially for many products stored in the frozen state.

Processing. The carbohydrate starch is rarely consumed in the raw state and in cooking is always gelatinized to some degree. For industrial purposes starches are submitted to many processes. Starch is often partially or almost wholly gelatinized or may be converted by heat or chemical treatment into dextrins for use in adhesive pastes, with the starch assuming a completely new form. Other treatments increase solubility, and hydrolysis with acids produces completely new products, including a variety of sugars.

Starch may be converted into sugars by the use of acids, and the sugars may be marketed as starch syrup, glucose syrup, or corn syrup; as glucose; and as commercial dextrose. Such sugars are useful in confectionery production.

Other uses of starch include production of ethyl alcohol by fermentation procedures and production of acetone and other products. Indeed, it is impossible to record all the hundreds of uses of starch in the science-based industries.

Alimentary pastes. *Pasta products.* Alimentary pastes include such products as macaroni, spaghetti, vermicelli, and noodles. Such products are often called pastas. Italy is regarded as the place of origin of macaroni products, and annual consumption in that country is as high as 30–35 kilograms (65–75 pounds) per person. Annual consumption is about 6.3 kilograms in France, 3.7 in the United States, and only 0.4 kilogram in the United Kingdom. Pasta is manufactured in a wide variety of sizes and shapes (see Figure 10), the commonest being long, narrow strands. The most slender type of strand, vermicelli, sometimes called *capelli d'angeli* ("angel's hair") in Italy, has a diameter ranging from 0.5 to 0.8 millimetre and is normally cut into lengths of about 250 millimetres and twisted into curls. Short-cut vermicelli (15–40 millimetres) is easy to manufacture and to dry. Spaghetti has a diameter of about 1.5–2.5 millimetres and is usually straight. Noodles are solid ribbons, about 0.8 millimetre thick, and in a variety of widths. Macaroni is the commonest type of alimentary paste; it is hollow and has a greater thickness than the others. It can be shaped in a variety of forms, such as long, short, large, small tubes, etc.

Annual consumption of pastas

Macaroni is now commercially produced in large facto-

Encyclopædia Britannica, Inc.

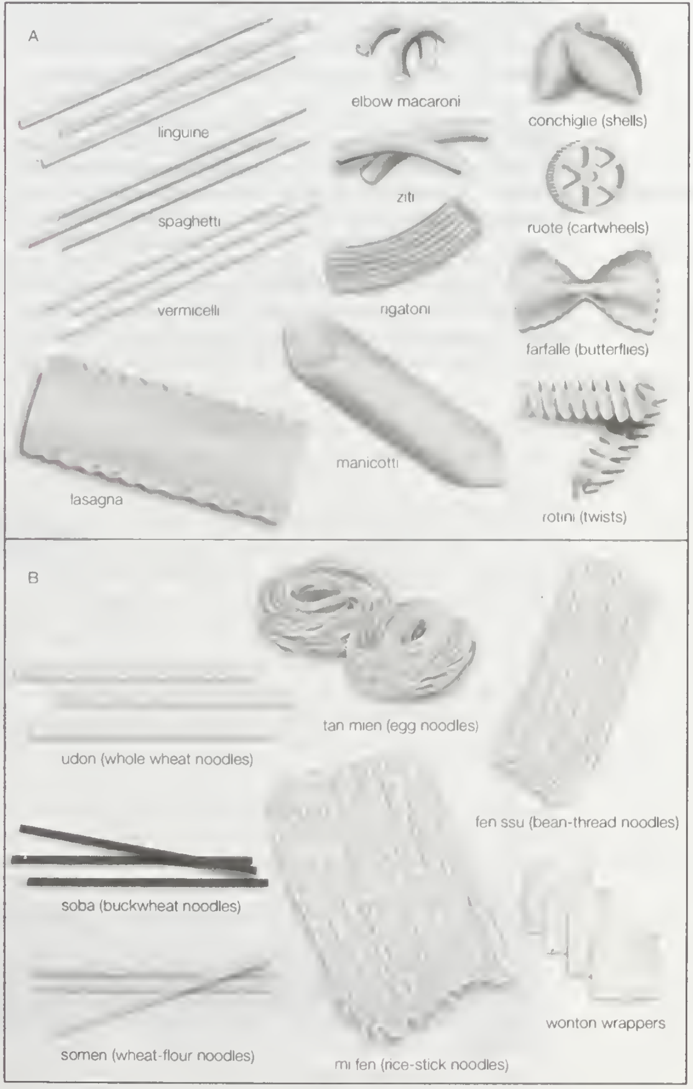


Figure 10: *Two types of alimentary pastes.*
(A) Italian-style pasta products; (B) Asian-style noodle products.

ries in Italy, North and South America, and other regions. Drying of the extruded paste is an important process, previously accomplished in Italy by sun drying.

Semolina. Semolina, not flour, is the form of cereal used, and various plain macaroni products are made by combining the correct form of semolina, from durum wheat, with water. Richer alimentary pastes are made with the addition of eggs in fresh, dried, or frozen form, and egg noodles are popular. In low-income families, alimentary pastes often provide the bulk of the calories in the diet. Macaroni products supply about 3,500 calories per kilogram and, although not themselves good sources of vitamins, are commonly cooked and consumed with butter, oil, cheese, and other items containing the needed vitamins.

The use of hard durum semolina contributes to good quality in macaroni and other alimentary paste products. The special mills involved use many breaks, and only a few reduction rolls, to produce as much clean semolina as possible. An efficient mill employing appropriate purifiers can produce as much as 65 percent semolina (together with a little flour). Before continuous processes for pasta production were introduced, a coarse semolina was valued. In modern production, semolina is dusted and freed from flour, and regularity in size is considered important for water absorption. Very fine semolina is not popular, and the preferred semolina usually has a moisture content of about 13 percent with less than 0.8 percent ash. Freedom from bran is desired to avoid the appearance of specks. The gluten in the semolina should be reasonably strong but not as elastic as that required for bread making. .

Pasta processing. In the early factories, batch mixing of semolina and water was followed by extrusion of the resulting paste through presses containing dies. In modern practice, the bulk of alimentary pastes is made by continuous processes.

Extrusion

The basic procedure for most macaroni products consists of adding water to a semolina made from suitable wheat to produce, in a short time, a plastic homogeneous mass of about 30 percent moisture. This mixture is extruded through special dies, under pressure, producing the desired size and shape, and is then dried (see Figure 11). There are many types of continuous paste processes adapted to the specific types of paste wanted and to the manufacturer's requirements. In the earlier days of the cottage industry, long-cut products such as spaghetti were spread evenly by hand on wooden dowels about an inch thick and over 50 inches long, and the filled sticks were then placed on racks for sun drying. Short-cut products were often scattered on wire mesh trays.

In modern automatic processing the objective is to dry the extruded product, containing 31 percent moisture, to a hard product of about 12 percent moisture, decreasing the possibility of the goods being affected by the growth of molds and yeast. If moisture is removed too rapidly, the dried product may tend to "check" or split. If moisture is removed too slowly, souring or mold growth may occur. Proper drying is therefore ensured by adjusting air circulation, temperature, and humidity. Drying procedures differ for long and short macaroni. In the continuous process, after a first hour in which a crust is formed to protect against mold infection, slow drying is practiced.

Testing. Cooking tests are used to ensure that the final product is satisfactory. Considerable research has been carried out to control factors tending to destroy the desirable yellow colour. Destruction of the colouring matter, a xanthophyll, can occur in mixing owing to excessive lipoxidase. Certain typès of durum wheat may possess a high degree of lipoxidase activity, and it is difficult to control or check this action. The addition of ascorbic acid has been suggested as a means to decrease the destruction of the semolina pigments in processing.

In the United States, alimentary paste goods, described as noodles, egg spaghetti, or egg macaroni, must contain 5.5 percent of the solids of egg in the final product. The eggs can be used in the form of frozen yolks, dried yolks, frozen whole eggs, dried whole eggs, or fresh whole eggs or yolks. Spray-dried egg yolks of good quality are now available.

Kellogg and Post

Breakfast cereals. *Origins.* The modern packaged breakfast-food industry owes its beginnings to an American religious sect, the Seventh-day Adventists, who wished to avoid consumption of animal foods. In the 1860s they organized the Western Health Reform Institute in Battle Creek, Mich., later renamed the Battle Creek Sanitarium. James Jackson of Dansville, N.Y., produced a cereal food by baking whole-meal dough in thin sheets, breaking and regrinding into small chunks, rebaking and regrinding. J.H. Kellogg of Battle Creek made biscuits about one-half inch thick from a dough mixture of wheatmeal, oatmeal, and cornmeal. The dough was baked until it was fairly dry and turning brown, and the product was ground and packed. A patient at the sanitarium, C.W. Post, saw the possibilities in such a product entirely apart from the

Encyclopædia Britannica, Inc.

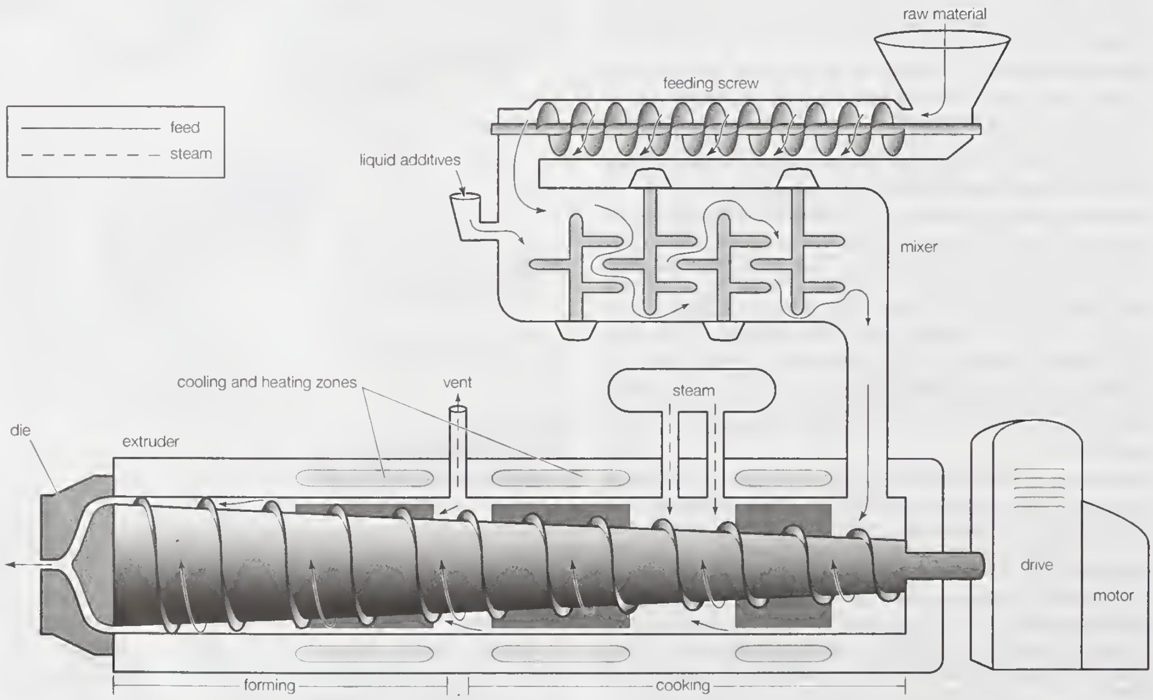


Figure 11: A high-temperature short-time extruder.

original conception of healthfulness and started a business. Kellogg's brother, W.K. Kellogg, did likewise, and the breakfast-food industry was launched, soon achieving mass sales of cereal products in flaked, granular, shredded, and puffed forms, with flavour obtained by roasting and the addition of sugar.

Types of breakfast cereal. Some breakfast cereals require cooking; others are packaged ready-to-eat. Roasted and rolled oatmeal, eaten as porridge, requires brief boiling. Cooking time of these processed cereals has been greatly reduced, and various "instant" forms are available.

Although cooked oatmeal porridge was formerly a standard breakfast food, the ready-to-eat cereals of various types are now the favourite breakfast-cereal foods.

The middlings produced in flour milling, essentially small pieces of endosperm free from bran and germ, are sold as farina and often consumed as a breakfast food in the United States. Farina is usually enriched with vitamins and minerals and may be flavoured. To reduce cooking time, 0.25 percent disodium phosphate may be added; some products require only one minute of boiling before serving.

Ready-to-eat cereals are available in a variety of forms and are normally consumed with milk and sometimes sugar. Flaked and toasted varieties are the most popular. During processing the starch is gelatinized, halting enzymatic reactions and thus ensuring product stability and good shelf life. The sugar content is dextrinized and caramelized by a roasting process. Roasting also ensures attractive crispness resulting from moisture reduction.

Flaked cereals. Wheat and rice flakes are manufactured, but most flaked breakfast foods are made from corn (maize), usually of the yellow type, broken down into grits and cooked under pressure with flavouring syrup consisting of sugar, nondiastatic malt, and other ingredients. Cooking is often accomplished in slowly rotating retorts under steam pressure.

After leaving the cooker, the lumps (containing about 33 percent water) are broken down by revolving reels and sent to driers. These are usually large tubes extending vertically, through several stories, with the wet product entering the top and encountering a current of hot air (65° C, or 150° F). Other types of driers consist of horizontal rotating cylinders with steam-heated pipes running horizontally. The drying process reduces moisture to about 20 percent, and the product is transferred to tempering bins for up to 24 hours, to even moisture distribution.

The product is next flaked by passing it between large steel cylinders (180–200 revolutions per minute), with the rolls cooled by internal water circulation. The cooked and rather soft flakes then proceed to rotating toasting ovens (normally gas-fired), where the flakes tumble through perforated drums. This treatment requires two to three minutes at 225° C (550° F). The product is dehydrated, toasted, and slightly blistered. After toasting it is cooled by circulating air, and at this stage enrichment by sprays may be carried out.

The manufacture of wheat flakes is similar to that of corn flakes. Special machinery separates the individual grains so that they can be flaked and finally toasted.

Shredded cereals. Shredded wheat, differing from other breakfast foods, is made from whole grains with the germ and bran retained and no flavour added. In its final form it is in tablets composed of shreds of cooked and toasted wheat. The wheat is cleaned and then boiled in water, often at atmospheric pressure. The grains reach a moisture content of 55 to 60 percent and require preliminary drying to about 50 percent. They are then placed in bins to condition them. The shredding process consists of passing the cooked and partially dried wheat to the shredding rolls, which are 150 to 200 millimetres (6 to 8 inches) in diameter and as wide as the finished tablet. On one pair of the rolls is a series of about 20 shallow corrugations running around the periphery; the surface of the other roll is smooth. The soft wheat is forced into the rolls under pressure and is cut into long shreds falling to a conveyor in such a way as to obtain superimposed shreds. These layers are cut into tablets by knives, and the tablets are transferred to baking pans. The pans pass to a revolving oven, with a baking temperature of approximately 260° C (500° F). After 10–15 minutes the outside of the product is dry and toasted, while the interior is still damp. The tablet is transferred either to another hot air oven or to a different section of the same oven, where it is dried at 120° C (250° F) for an additional 30 minutes and then cooled and packed.

Granular cereals. Granular types are made by very different processes from the others. The first step is production of a stiff dough from wheat, malted barley flour, salt, dry yeast, and water. After mixing, fermentation proceeds for about five hours. The dough is then formed into large loaves and transferred directly to the oven. Baking requires about two hours at 205° C (400° F). The baked loaves are fragmented and the product is then thoroughly dried. Grinding by corrugated rolls follows, and the product is sieved to standard size. Very finely ground pieces are added to subsequent dough batches.

Puffed cereals. Early in the 20th century an American patent was taken out for the preparation of puffed wheat and rice. Puffed oats and corn are now also produced. The principle of the puffing process is heating the cereal, and sometimes other vegetable products, in a pressure chamber to a pressure of 7 to 14 kilograms per square centimetre (100 to 200 pounds per square inch), then instantaneously releasing this pressure by suddenly opening the chamber, or puffing gun. Expansion of the water vapour occurs when the pressure is suddenly released, blowing up the grains or cereal pellets to several times their original size (8-fold to 16-fold for wheat, 6-fold to 8-fold for rice). The final product is toasted to a moisture content of about 3 percent to achieve desired crispness. In processing wheat, a preliminary step may be applied to free the grain from much of its bran coatings.

The puffing gun

Rice is usually parboiled, pearled, and cooked with sugar syrup, dried to about 25 to 30 percent moisture in rotating louvre dryers, binned, and toasted and puffed. In puffing of mixed cereal products it is necessary to start with a stiff dough containing sugar, salt, and sometimes oil, and this mixture is then cooked. The dough is pelleted by extrusion through dies and dried to attain a suitable condition for the final puffing process.

Enrichment. Enrichment of breakfast cereals with minerals, and especially with vitamins, is now common practice. In many of the manufacturing processes employed in breakfast-food production, considerable vitamin destruction occurs. The various heat treatments involved may destroy 90 percent of the original B_1 content of the cereal, especially in flaked and puffed products. On the other hand, a proportion of the somewhat harmful phytic acid in cereals, interfering with absorption by the body of calcium, is also destroyed; and enrichment of the products with vitamin B_1, and sometimes other components of the vitamin B complex, is not difficult to perform after the various cooking operations have been completed.

Sweeteners. Various types of sweeteners are made directly from starch. Glucose products made by starch conversion differ in composition and in sweetness according to whether conversion is effected by acid or by enzymes. Enzyme-produced glucose is higher in dextrose and maltose content than acid-produced glucose, which normally is higher in dextrin. Sucrose is a more powerful sweetening agent than dextrose, but glucose syrup made by enzymatic treatment usually has twice the sweetening power of that produced by acid action.

Corn syrup

In the production of starch separated by the wet milling of corn, one stream is normally used to produce starch, and the other stream is converted into corn syrup by heating the starch slurry in pressure tanks with acid or enzymes and following with refining processes. If the process of hydrolyzing starch is completed, the resulting product is glucose. Often the treatment is not carried to completion, and a series of dextrins and reversion products is produced. If full conversion is required, the treatment usually employs acids to liquefy, followed by saccharifying enzymes to complete the change to dextrose. Modern syrups and crystalline dextrose are made by continuous processes. The degree of conversion of the starch into the sugar dextrose is expressed as D.E. (dextrose equivalents), and confectionery syrups have a D.E. of about 36 to 55, while the fuller conversion of products with D.E. of 96 to 99 can be made for the production of almost pure glucose or dextrose, used in many food products.

Sweeteners in the form of syrups are largely used in cake and confectionery products and also, especially in the United States, in bread manufacture. American bread is distinctly sweeter than normal European bread because of the fats and sweeteners used, and the loaves are larger per unit of weight than in the United Kingdom and most European countries.

In the United States the baking industry uses more than one-half of the dextrose and 10 percent of the corn syrup produced. Makers of cookies (biscuits) and breakfast foods also use large amounts of sweeteners. Confectioners in Europe use syrups of many types but not as widely as in the United States. (D.W.K.-J./Ed.)

Edible fats and oils

The oil and fat products used for edible purposes can be divided into two distinct classes: liquid oils, such as olive oil, peanut oil, soybean oil, or sunflower oil; and plastic fats, such as lard, shortening, butter, and margarine. The physical nature of the fatty material is unimportant for some uses, but the consistency is a matter of consequence for other products. As a dressing on green salads, for example, a liquid oil is used to provide a coating on the ingredients; a plastic fat such as lard or butter would be unsuitable. Spreads for bread, foods that require a highly developed dough structure, or icings and fillings with a plastic structure require plastic fats rather than liquid oils.

GENERAL METHODS OF EXTRACTION

The raw materials for the fat and oil industry are animal by-products from the slaughter of cattle, hogs, and sheep; fatty fish and marine mammals; a few fleshy fruits (palm and olive); and various oilseeds. Most oilseeds are grown specifically for processing to oils and protein meals, but several important vegetable oils are obtained from by-product raw materials. Cottonseed is a by-product of cotton grown for fibre, and corn oil is obtained from the corn germ that accumulates from the corn-milling industry, whose primary products are corn grits, starch, and syrup.

Fats may be recovered from oil-bearing tissues by three general methods, with varying degrees of mechanical simplicity: (1) rendering, (2) pressing with mechanical presses, and (3) extracting with volatile solvents.

Rendering. *Fruits and seeds.* The crudest method of rendering oil from oleaginous fruits, still practiced in some countries, consists of heaping them in piles, exposing them to the sun, and collecting the oil that exudes. In a somewhat improved form, this process is used in the preparation of palm oil; the fresh palm fruits are boiled in water, and the oil is skimmed from the surface. Such processes can be used only with seeds or fruits (such as olive and palm) that contain large quantities of easily released fatty matter.

Animal fats. The rendering process is applied on a large scale to the production of animal fats such as tallow, lard, bone fat, and whale oil. It consists of cutting or chopping the fatty tissue into small pieces that are boiled in open vats or cooked in steam digesters. The fat, gradually liberated from the cells, floats to the surface of the water, where it is collected by skimming. The membranous matter (greaves) is separated from the aqueous (gluey) phase by pressing in hydraulic or screw presses; additional fat is thereby obtained. The residue is used for animal feed or fertilizer. Several centrifugal separation processes were developed in the 1960s. Cells of the fatty tissues are ruptured in special disintegrators under close temperature control. The protein tissue is separated from the liquid phase in a desludging type of centrifuge, following which a second centrifuge separates the fat from the aqueous protein layer. Compared with conventional rendering, the centrifugal methods provide a higher yield of better-quality fat, and the separated protein has potential as an edible meat product.

Pressing. *Pressing processes.* With many oil-bearing seeds and nuts, rendering will not liberate the oil from the cellular structures in which it is held (see Figure 12). In these cases the cell walls are broken by grinding, flaking, rolling, or pressing under high pressures to liberate the oil. The general sequence of modern operations in pressing oilseeds and nuts is as follows: (1) the seeds are passed over magnetic separators to remove any stray bits of metal; (2) if necessary, the shells or hulls are removed; (3) the kernels or meats are converted to coarse meal by grinding them between grooved rollers or with special types of hammer mills; and (4) they are pressed in hydraulic or screw presses with or without preliminary heating, depending on the type of oil-bearing material and the quality of oil desired. Oil expressed without heating contains the least amount of impurities and is often of edible quality without refining or further processing. Such oils are known as cold-drawn, cold-pressed, or virgin oils. Pressing the coarse meal while it is heated removes more oil and also greater quantities of nonglyceride impurities such as phospholipids, colour bodies, and unsaponifiable matter. Such oil is more highly coloured than cold-pressed oils. Residual meals are concentrated sources of high-quality protein and are generally used in animal feeds.

Quality variations in pressed oils

Encyclopædia Britannica, Inc.

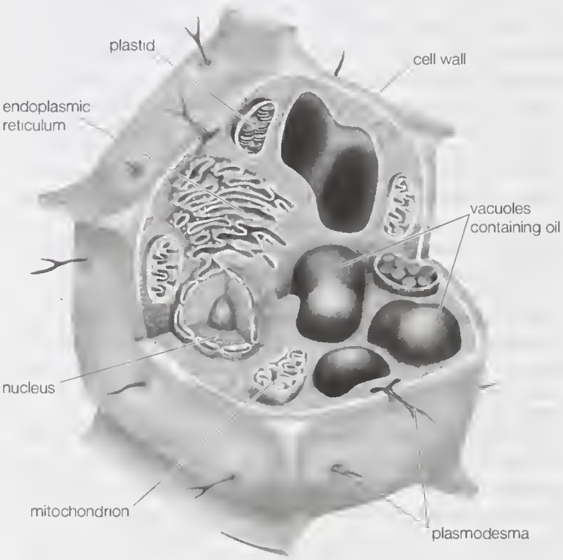


Figure 12: Some of the structures of an oilseed cell, including oil-containing vacuoles.

Pressing machines. Many different mechanical devices have been used for pressing. The Romans developed a screw press, described by Pliny, for the production of olive oil. Centuries ago, the Chinese employed the same series of operations followed in modern pressing mills—namely, bruising or grinding the seeds in stone mills, heating the meal in open pans, and then pressing out the oil in a wedge press. The Dutch, or stamper, press invented in the 17th century was used almost exclusively in Europe for pressing oilseeds until the early part of the 19th century, when the hydraulic press was developed. The yield of oil from the hydraulic press was considerably higher than that from earlier processing methods because of the much higher applied pressures. In open presses, the ground seed material was confined in cloths of human hair or, less commonly, camel hair. Pressures on the cake varied from approximately 70 to 140 kilograms per square centimetre (1,000 to 2,000 pounds per square inch), and in the closed-type press, in which the oil-containing material was confined in a strong perforated steel cage during the pressing operation, pressures of approximately 400 kilograms per square centimetre or more were attained. Under ideal conditions the oil content of the hydraulic-press cake can be reduced to about 3 percent, but in practical operation a 5 percent level is average. The modern screw press replaced many of the hydraulic presses because it is a continuous process, has greater capacity, requires less labour, and will generally remove more oil. As ground seed is fed continuously into the mechanical press, a worm screw increases the pressure progressively as the material moves through a slotted barrel. Pressures from 700 to 2,100 kilograms per square centimetre are attained, and the oil is squeezed out through the slots, leaving a cake containing 3 to 3.5 percent oil under optimum processing and 4 to 5 percent oil under average conditions.

Hydrocarbon solvents

Solvent extraction. *Processes.* Cakes obtained by pressing operations still retain 3 to 15 percent of residual oil. When the value of the oil is considerably greater as oil than as a part of the meal, it is desirable to obtain more complete extraction with solvents. Modern commercial methods of solvent extraction use volatile purified hydrocarbons, especially the various grades of petroleum benzin (commonly known as petroleum ether, commercial hexane, or heptane). In large-scale operations, solvent extraction is a more economical means of recovering oil than is mechanical pressing. In the United States and increasingly in Europe, there are many instances of simple petroleum benzin extraction of seeds, mainly soybeans. For seeds or nuts containing a higher oil content than soybeans it became customary to press the material in screw presses to remove a large proportion of the oil before extraction. Since this prepressing also ruptures the cellular structures of oil-bearing materials, most of the residual oil is easily removed with solvents.

A typical extraction system consists of (1) cleaning to remove tramp iron, dirt, foreign weed seeds, and stones, (2) removing hulls or cortex in cracking, aspirating, or screening operations, (3) cracking or rough grinding the kernels, meats, or prepressed cake, (4) steaming (tempering or cooking) of the meats, (5) flaking the small pieces between smooth flaking rolls, (6) extracting the oil with solvent, (7) separating the meal, or marc, from the oil-solvent solution, called miscella, and (8) removing the solvent from both the miscella and the marc. The marc may be toasted or pelletized, or both, for use in animal feeds. Most extracted meals contain less than 1 percent of residual oil. The amount varies depending on the amount of prepressing, the type of material being extracted, and the efficiency of the extracting system.

Extractors. Solvent extraction was first practiced in Europe, using batch extractors for the recovery of additional oil from the residues obtained from mechanical pressing. The greater efficiency of solvent extraction encouraged direct application to oilseeds, and the batch extractor gradually gave way to continuous units in which fresh flakes are added continuously and subjected to a counterflow of solvent. One of the earliest continuous extractors, and a type still considered to be one of the best, was the Bollman or Hansa-Mühle unit from Germany, in which solvent percolates through oilseed flakes contained in perforated baskets moving on an endless chain. After the extraction cycle is complete, the baskets of extracted flakes are dumped automatically and then refilled with fresh flakes to initiate another cycle. Many extractor designs have been proposed, but only a few have found wide acceptance. In the DeSmet extractor, popular in Europe and in a number of developing countries, a bed of flakes on an endless horizontal traveling belt is extracted by solvent percolation. The Blaw-Knox Rotocell has become the most popular extractor in the huge American soybean industry. The flakes are conveyed into wedge-shaped segments of a large cylindrical vessel. Solvent percolating through the cells falls into the bottom of the extractor housing, where it is picked up by a series of pumps and recirculated countercurrent to the flakes.

PROCESSING OF EXTRACTED OIL

The extent of processing applied to fats depends on their source, quality, and ultimate use. Many fats are used for edible purposes after only a single processing step—*i.e.,* clarification by settling or filtering. Most cold-pressed oils (for example, cold-pressed olive, peanut, and some coconut and sunflower oils) can be used in food products without further processing. Tremendous quantities of butter and lard are used without special treatment after churning or rendering. The growing demand for bland-tasting and stable salad oils and shortening, however, has led to extensive processing techniques.

Refining. The nonglyceride components contribute practically all the colour and flavour to fats. In addition, such materials as the free fatty acids, waxes, colour bodies, mucilaginous materials, phospholipids, carotenoids, and gossypol (a yellow pigment found only in cottonseed oil) contribute other undesirable properties in fats used for edible and, to some extent, industrial purposes.

Batch and continuous refining

Alkali refining. Many of these can be removed by treating fats at 40° to 85° C (104° to 185° F) with an aqueous solution of caustic soda (sodium hydroxide) or soda ash (sodium carbonate). The refining may be done in a tank (in which case it is called batch or tank refining) or in a continuous system. In batch refining, the aqueous emulsion of soaps formed from free fatty acids, along with other impurities (soapstock), settles to the bottom and is drawn off. In the continuous system the emulsion is separated with centrifuges. After the fat has been refined, it is usually washed with water to remove traces of alkali and soapstock. Oils that have been refined with soda ash or ammonia generally require a light re-refining with caustic soda to improve colour. After water washing, the oil may be dried by heating in a vacuum or by filtering through a dry filter-aid material. The refined oil may be used for industrial purposes or may be processed further to edible oils. Usually, the refined oils are neutral (*i.e.,* neither acidic nor alkaline), free of material that separates on heating (break material), lighter in colour, less viscous, and more susceptible to rancidity.

Water refining. Water refining, usually called degumming, consists of treating the natural oil with a small amount of water, followed by centrifugal separation. The process is applied to many oils that contain phospholipids in significant amounts. Since the separated phospholipids are rather waxy or gummy solids, the term degumming was quite naturally applied to the separation. The separated phospholipid emulsion layer from oils such as corn (maize) and soybean oils may be dried (commercially, these products are called lecithin) and used as emulsifiers in such products as margarine, chocolate products, and emulsion paints. The degumming of crude soybean oil, which has an average phospholipid content of 1.8 percent, provides the primary source of commercial lecithin. To obtain products of lighter colour, hydrogen peroxide may be added as a bleaching agent during the drying of lecithin. The degummed oil may be used directly in industrial applications, such as in paints or alkyd resins, or refined with alkalies for ultimate edible consumption.

Bleaching. If further colour removal is desired, the fat may be treated with various bleaching agents. Heated oils are treated with fuller's earth (a natural earthy material that will decolorize oils), activated carbon, or activated clays. Many impurities, including chlorophyll and carotenoid pigments, are adsorbed onto such agents and removed by filtration. Bleaching often reduces the resistance of oils to rancidity, because some natural antioxidants are removed together with impurities. When many oils are heated to more than 175° C (347° F), a phenomenon known as heat bleaching takes place. Apparently the heat decomposes some pigments, such as the carotenoids, and converts them to colourless materials.

Removing traces of wax

Destearinating or winterizing. It is often desirable to remove the traces of waxes (*e.g.,* cuticle wax from seed coats) and the higher-melting glycerides from fats. Waxes can generally be removed by rapid chilling and filtering. Separation of high-melting glycerides, or stearine, usually requires very slow cooling in order to form crystals that are large enough to be removed by filtration or centrifuging. Thus linseed oil may be winterized to remove traces of waxes that otherwise interfere with its use in paints and varnishes. Stearine may be removed from fish oils in order to separate the solid glycerides that would detract from its use in paints and alkyd resins. At the same time, fish stearine is more suitable than whole oil for edible purposes. Cottonseed and peanut oils may be destearinated to produce salad oils that remain liquid at low temperatures. Tallows and other animal fats may be destearinated for simultaneous production of hard fats (high in stearic acid content for special uses such as in making candles) and of liquid oil called oleo oil.

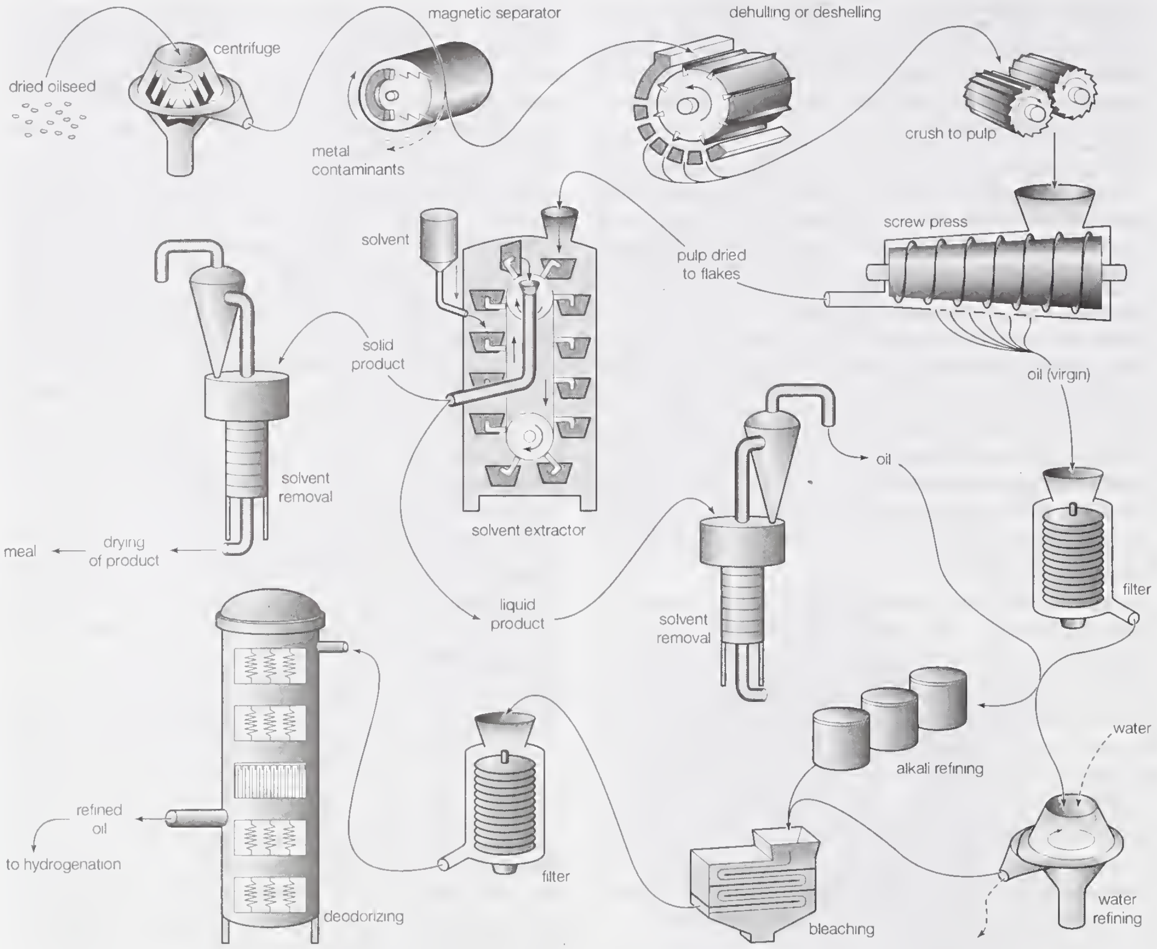


Figure 13: Essential steps in the extracting and refining of edible oil from oilseeds.
Encyclopædia Britannica, Inc.

Hydrogenation. For many edible purposes and for some commercial applications it is desirable to produce solid fats. Many shortenings and margarines contain hydrogenated (hardened) oils as their major ingredients. The development of margarine and shortening products resulted from the invention of a successful method for converting low-melting unsaturated fatty acids and glycerides to higher-melting saturated products. The process consists of the addition of hydrogen in the presence of a catalyst to the double (unsaturated) bonds. Thus oleic or linoleic acid (or their acid radicals in glycerides), which are normally liquid at room temperature, can be converted to stearic acid or the acid radical by the addition of hydrogen.

Limited use was made of this hydrogenation technology in Europe; the greatest potential use for the process lay in the United States, where a vast production of cottonseed oil, a by-product of the Southern cotton industry, awaited developments that would permit its conversion to a plastic fat. The hardening of cottonseed oil in the early 1900s gave birth to the shortening industry. Practical hydrogenation then spread to all countries where margarines and shortenings are produced from liquid oils.

Hydrogenation reactions. In commercial practice, hydrogenation is usually carried out with vigorous agitation or hydrogen dispersion with a narrow range of catalyst concentration (about 0.05 to 0.10 percent of finely divided nickel suspended on kieselguhr, or diatomaceous earth) in a steel pressure-reaction vessel. The ordinary ranges of temperature and pressure are from 100° to 200° C (212° to 392° F) and from atmospheric pressure to 42 kilograms per square centimetre, respectively. These conditions can be controlled to make the hydrogenation reaction somewhat selective—*i.e.,* to add hydrogen to the linolenic (three double bonds) and linoleic (two double bonds) acid radicals before adding to the oleic (one double bond) acid radicals. The most unsaturated fatty acid groups are most easily hydrogenated and thus react first with the hydrogen if conditions are right. Copper-containing catalysts are especially selective in the hydrogenation of vegetable oils. If very hard fats with low amounts of unsaturation are desired and selectivity is unimportant, higher temperatures and pressures are employed to shorten the reaction time and to use partially spent catalyst that would otherwise be wasted. After hydrogenation, the hot oil is filtered to remove the metallic catalyst for either reuse or recovery.

Isomerization reactions. During the catalytic treatment another reaction also takes place—isomerization (rearrangement of the molecular structure) of unsaturated fatty acid radicals to form isooleic, isolinoleic, and similar groups. Because these isomers have higher melting points than do the natural acids, they contribute to the hardening effect. The unsaturation of natural oils has the *cis* configuration, in which hydrogen atoms lie on one side of a plane cutting through the double bond and alkyl groups lie on the other side. During hydrogenation some of the unsaturation is converted to the *trans* configuration, with like groups on opposite sides of the plane. The *trans* isomers are much higher melting than the natural *cis* form. Simultaneously with the change of some of the unsaturation to the *trans* configuration there is a migration of double bonds along the chain. Thus isomers of oleic acid may be formed with the double bond in any position from carbon atom 2 to carbon atom 17. Many of these isomerized acids are higher melting than the natural oleic acid. Infrared analysis is useful for quantitative measurement of changes occurring during hydrogenation.

Rearrangement of molecular structure

Deodorization. Odourless and tasteless fats first came into high demand as ingredients for the manufacture of margarine, a product designed to duplicate the flavour and texture of butter. Most fats, even after refining, have characteristic flavours and odours, and vegetable fats especially have a relatively strong taste that is foreign to that of butter. The deodorization process consists of blowing steam through heated fat held under a high vacuum. Small quantities of volatile components, responsible for tastes and odours, distill, leaving a neutral, virtually odourless fat that is suitable for the manufacture of bland shortening or delicately flavoured margarine. Originally, deodorization was a batch process, but increasingly, continuous systems are being used in which hot fat flows through an evacuated column countercurrent to the upward passage of steam. In Europe, a deodorization temperature of 175°–205° C (347°–401° F) is common, but in the United States, higher temperatures of 235°–250° C (455°–483° F) are usually employed. About 0.01 percent of citric acid is commonly added to deodorized oils to inactivate trace-metal contaminants such as soluble iron or copper compounds that otherwise would promote oxidation and the development of rancidity.

Characteristics of cooking oils

Olive oil is invariably marketed in undeodorized form. The natural flavour is an important asset, and olive oil, as is true of butter, commands a premium in the market because of its distinctive and prized flavour. The common cooking oils of Asia—soybean, rapeseed, peanut, sesame, and coconut—are consumed in their crude form as expressed from oilseeds. In contrast, deodorized oils are in particular demand in the United States and Europe. For many years the only important vegetable oil consumed in the United States was cottonseed oil, which in its crude form has such a strong and unpleasant flavour that further processing was an absolute necessity in order to render it suitable for consumption. Because of widespread sale of neutral-flavoured cottonseed oil products over many years, a general preference was developed for odourless and tasteless fats.

Another reason for the practice of deodorizing edible oils in Europe and America relates to differences in oil quality by Western and Eastern extraction techniques. In China and Southeast Asia, edible oils have been produced principally by small, relatively crude equipment. The yield of oil is relatively low, and a minimum amount of nonglyceride substances is expressed from the seed, with the result that the flavour of the oil is fairly mild. In Europe and the United States, oil extraction is carried out in large factories that operate on an extremely competitive basis. Very-high-pressure expression or solvent extraction is used, and in order to improve yields the seeds are heat-treated prior to extraction. Oils obtained in high yield under such conditions are stronger in flavour than oils prepared by low-pressure expression, and the refining and deodorizing steps are required to improve palatability. The improvement in yields more than compensates for the added costs of refining and deodorizing.

When fats are hydrogenated for manufacture of margarine and shortening, they develop a characteristic sweet, but rather unpleasant, "hydrogenation odour" that must be removed from edible fats by deodorization.

(A.R.B./M.W.F./Ed.)

Bakery products

Baking, a dry-heat cooking process, is probably the oldest cooking method. Bakery products are usually prepared from flour or meal derived from some form of grain. Bread, already a common staple in prehistoric times, provides many nutrients in the human diet.

HISTORY

The earliest processing of cereal grains probably involved parching or dry roasting of collected grain seeds. Flavour, texture, and digestibility were later improved by cooking whole or broken grains with water, forming gruel or porridge. It was a short step to the baking of a layer of viscous gruel on a hot stone, producing primitive flat bread. More sophisticated versions of flat bread include the Mexican tortilla, made of processed corn, and the chapati of India, usually made of wheat.

Yeast leavening

Baking techniques improved with the development of an enclosed baking utensil and then of ovens, making possible thicker baked cakes or loaves. The phenomenon of fermentation, with the resultant lightening of the loaf structure and development of appealing flavours, was probably first observed when doughs or gruels, held for several hours before baking, exhibited spoilage caused by yeasts. Some of the effects of the microbiologically induced changes were regarded as desirable, and a gradual acquisition of control over the process led to traditional methods for making leavened bread loaves. Early baked products were made of mixed seeds with a predominance of barley, but wheat flour, because of its superior response to fermentation, eventually became the preferred cereal among the various cultural groups sufficiently advanced in culinary techniques to make leavened bread.

Brewing and baking were closely connected in early civilizations. Fermentation of a thick gruel resulted in a dough suitable for baking; a thinner mash produced a kind of beer. Both techniques required knowledge of the "mysteries" of fermentation and a supply of grain. Increasing knowledge and experience taught the artisans in the baking and brewing trades that barley was best suited to brewing, while wheat was best for baking.

By 2600 BC the Egyptians, credited with the first intentional use of leavening, were making bread by methods similar in principle to those of today. They maintained stocks of sour dough, a crude culture of desirable fermentation organisms, and used portions of this material to inoculate fresh doughs. With doughs made by mixing flour, water, salt, and leaven, the Egyptian baking industry eventually developed more than 50 varieties of bread, varying the shape and using such flavouring materials as poppyseed, sesame, and camphor. Samples found in tombs are flatter and coarser than modern bread.

The Egyptians developed the first ovens. The earliest known examples are cylindrical vessels made of baked Nile clay, tapered at the top to give a cone shape and divided inside by a horizontal shelflike partition. The lower section is the firebox, the upper section is the baking chamber. The pieces of dough were placed in the baking chamber through a hole provided in the top.

In the first two or three centuries after the founding of Rome, baking remained a domestic skill with few changes in equipment or processing methods. According to Pliny the Elder, there were no bakers in Rome until the middle of the 2nd century BC. As well-to-do families increased, women wishing to avoid frequent and tedious bread making began to patronize professional bakers, usually freed slaves. Loaves molded by hand into a spheroidal shape, generally weighing about a pound, were baked in a beehive-shaped oven fired by wood. *Panis artopticius* was a variety cooked on a spit, *panis testuatis* in an earthen vessel.

Although Roman professional bakers introduced technological improvements, many were of minor importance, and some were essentially reintroductions of earlier developments. The first mechanical dough mixer, attributed to Marcus Virgilius Euryasaces, a freed slave of Greek origin, consisted of a large stone basin in which wooden paddles, powered by a horse or donkey walking in circles, kneaded the dough mixture of flour, leaven, and water.

Bakers' guilds

Guilds formed by the miller-bakers of Rome became institutionalized. During the 2nd century AD, under the Flavians, they were organized into a "college" with work rules and regulations prescribed by government officials. The trade eventually became obligatory and hereditary, and the baker became a kind of civil servant with limited freedom of action.

During the early Middle Ages, baking technology advances of preceding centuries disappeared, and bakers reverted to mechanical devices used by the ancient Egyptians and to more backward practices. But in the later Middle Ages the institution of guilds was revived and expanded. Several years of apprenticeship were necessary before an applicant was admitted to the guild; often an intermediate status as journeyman intervened between apprenticeship and full membership (master). The rise of the bakers' guilds reflected significant advances in technique. A 13th-century French writer named 20 varieties of bread varying in shape, flavourings, preparation method, and quality of the meal used. Guild regulations strictly governed size and quality. But outside the cities bread was usually baked in the home. In medieval England rye was the main ingredient of bread consumed by the poor; it was frequently diluted with meal made from other cereals or leguminous seeds. Not until about 1865 did the cost of white bread in England drop below brown bread.

At that time improvements in baking technology began to accelerate rapidly, owing to the higher level of technology generally. Ingredients of greater purity and improved functional qualities were developed, along with equipment reducing the need for individual skill and eliminating hand manipulation of bread doughs. Automation of mixing, transferring, shaping, fermentation, and baking processes began to replace batch processing with continuous operations. The enrichment of bread and other bakery foods with vitamins and minerals was a major accomplishment of the mid-20th-century baking industry.

INGREDIENTS

Flour, water, and leavening agents are the ingredients primarily responsible for the characteristic appearance, texture, and flavour of most bakery products. Eggs, milk, salt, shortening, and sugar are effective in modifying these qualities, and various minor ingredients may also be used.

Flour. Wheat flour is unique among cereal flours in that, when mixed with water in the correct proportions, its protein component forms an elastic network capable of holding gas and developing a firm spongy structure when baked. The proteinaceous substances contributing these properties are known collectively as gluten. The suitability of a flour for a given purpose is determined by the type and amount of its gluten content. These characteristics are controlled by the genetic constitution and growing conditions of the wheat from which the flour was milled, as well as the milling treatment applied.

Gluten

Low-protein, soft-wheat flour is appropriate for cakes, pie crusts, cookies (sweet biscuits), and other products not requiring great expansion and elastic structure. High-protein, hard-wheat flour is adapted to bread, hard rolls, soda crackers, and Danish pastry, all requiring elastic dough and often expanded to low densities by the leavening action.

Leavening agents. Pie doughs and similar products are usually unleavened, but most bakery products are leavened, or aerated, by gas bubbles developed naturally or folded in. Leavening may result from yeast or bacterial fermentation, from chemical reactions, or from the distribution in the batter of atmospheric or injected gases.

Yeast. All commercial breads, except salt-rising types and some rye bread, are leavened with bakers' yeast, composed of living cells of the yeast strain *Saccharomyces cerevisiae.* A typical yeast addition level might be 2 percent of the dough weight. Bakeries receive yeast in the form of compressed cakes containing about 70 percent water or as dry granules containing about 8 percent water. Dry yeast, more resistant to storage deterioration than compressed yeast, requires rehydration before it is added to the other ingredients. "Cream" yeast, a commercial variety of bakers' yeast made into a fluid by the addition of extra water, is more convenient to dispense and mix than compressed yeast, but it also has a shorter storage life and requires additional equipment for handling.

Bakers' yeast performs its leavening function by fermenting such sugars as glucose, fructose, maltose, and sucrose. It cannot use lactose, the predominant sugar of milk, or certain other carbohydrates. The principal products of fermentation are carbon dioxide, the leavening agent, and ethanol, an important component of the aroma of freshly baked bread. Other yeast activity products also flavour the baked product and change the dough's physical properties.

Conversion of sugars to carbon dioxide and ethanol

The rate at which gas is evolved by yeast during the various stages of dough preparation is important to the success of bread manufacture. Gas production is partially governed by the rate at which fermentable carbohydrates become available to the yeast. The sugars naturally present in the flour and the initial stock of added sugar are rapidly exhausted. A relatively quiescent period follows, during which the yeast cells become adapted to the use of maltose, a sugar constantly being produced in the dough by the action of diastatic enzymes on starch. The rate of yeast activity is also governed by temperature and osmotic pressure, the latter primarily a function of the water content and salt concentration.

Baking soda. Layer cakes, cookies (sweet biscuits), biscuits, and many other bakery products are leavened by carbon dioxide from added sodium bicarbonate (baking soda). Added without offsetting amounts of an acidic substance, sodium bicarbonate tends to make dough alkaline, causing flavour deterioration and discoloration and slowing carbon dioxide release. Addition of an acid-reacting substance promotes vigorous gas evolution and maintains dough acidity within a favourable range.

Carbon dioxide produced from sodium bicarbonate is initially in dissolved or combined form. The rate of gas release affects the size of the bubbles produced in the dough, consequently influencing the grain, volume, and texture of the finished product. Much research has been devoted to the development of leavening acids capable of maintaining the rate of gas release within the desired range. Acids such as acetic, from vinegar, or lactic, from sour milk, usually act too quickly; satisfactory compounds include cream of tartar (potassium acid tartrate), sodium aluminum sulfate (alum), sodium acid pyrophosphate, and various forms of calcium phosphate.

Baking powder. Instead of adding soda and leavening acids separately, most commercial bakeries and domestic bakers use baking powder, a mixture of soda and acids in appropriate amounts and with such added diluents as starch, simplifying measuring and improving stability. The end products of baking-powder reaction are carbon dioxide and some blandly flavoured harmless salts. All baking powders meeting basic standards have virtually identical amounts of available carbon dioxide, differing only in reaction time. Most commercial baking powders are of the double-acting type, giving off a small amount of available carbon dioxide during the mixing and makeup stages, then remaining relatively inert until baking raises the batter temperature. This type of action eliminates excessive loss of leavening gas, which may occur in batter left in an unbaked condition for long periods.

Double-acting baking powder

Entrapped air and vapour. Angel food cakes, sponge cakes, and similar products are customarily prepared without either yeast or chemical leaveners. Instead, they are leavened by air entrapped in the product through vigorous beating. This method requires a readily foaming ingredient capable of retaining the air bubbles, such as egg whites. To produce a cake of fine and uniform internal structure, the pockets of air folded in during beating are rapidly subdivided into small bubbles with such mixing utensils as wire whips, or whisks.

The vaporization of volatile fluids (*e.g.,* ethanol) under the influence of oven heat can have a leavening effect. Water-vapour pressure, too low to be significant at normal temperatures, exerts substantial pressure on the interior walls of bubbles already formed by other means as the interior of the loaf or cake approaches the boiling point. The expansion of such puff pastry as used for napoleons (rich desserts of puff pastry layers and whipped cream or custard) and *vol-au-vents* (puff pastry shells filled with meat, fowl, fish, or other mixtures) is entirely due to water-vapour pressure.

Shortening. Fats and oils are essential ingredients in nearly all bakery products. Shortenings have a tenderizing effect in the finished product and often aid in the manipulation of doughs. In addition to modifying the mouth feel or texture, they often add flavour of their own and tend to round off harsh notes in some of the spice flavours.

The common fats used in bakery products are lard, beef fats, and hydrogenated vegetable oils. Butter is used in some premium and specialty products as a texturizer and to add flavour, but its high cost precludes extensive use. Cottonseed oil and soybean oil are the most common processed vegetable oils used. Corn, peanut, and coconut oils are used to a limited extent; fats occurring in other ingredients, such as egg yolks, chocolate, and nut butters, can have a shortening effect if the ingredients are present in sufficient quantity.

Breads and rolls often contain only 1 or 2 percent shortening; cakes will have 10 to 20 percent; Danish pastries prepared according to the authentic formula may have about 30 percent; pie crusts may contain even more. High usage levels require those shortenings that melt above room temperature; butter and liquid shortenings, with their lower melting point, tend to leak from the product.

Commercial shortenings may include antioxidants, to retard rancidity, and emulsifiers, to improve the shortening effect. Colours and flavours simulating butter may also be added. Margarines, emulsions of fat, water, milk solids, and salt, are popular bakery ingredients.

Fats of any kind have a destructive effect on meringues and other protein-based foams; small traces of oil left on the mixing utensils can deflate an angel food cake to unacceptably high density.

The use of water in dough

Liquids. Water is the liquid most commonly added to doughs. Milk is usually added to commercial preparations in dried form, and any moisture added in the form of eggs and butter is usually minimal. Water is not merely a diluent or inert constituent; it affects every aspect of the finished product, and careful adjustment of the amount of liquid is essential to make the dough or batter adaptable to the processing method. If dough is too wet it will stick to equipment and have poor response to shaping and transfer operations; if too dry, it will not shape or leaven properly.

Water hydrates gluten, permitting it to aggregate in the form of a spongy cellular network, the structural basis of most bakery products. It provides a medium in which yeast can metabolize sugars to form carbon dioxide and flavouring components and allows diffusion of nutrients and metabolites throughout the mass. Water is an indispensable component of the baking-powder reaction, and it allows starch to gelatinize during baking and prevents interior browning of bakery products.

Water impurities affect dough properties. Water preferred for baking is usually of medium hardness (50 to 100 parts per million) with a neutral pH (degree of acidity), or slightly acid (low pH). Water that is too soft can result in sticky doughs, while very hard water may retard dough expansion by toughening the gluten (calcium ions, particularly, promote cross-linking of gluten protein molecules). Water sufficiently alkaline to raise the dough pH may have a deleterious effect on fermentation and on flour enzymes. Although strongly flavoured contaminants may affect the acceptability of the finished product, chlorides and fluorides at concentrations usually found in water supplies have little influence on bread doughs.

Eggs. The differences between yolks and whites must be recognized in considering the effect of eggs on bakery products. Yolks contain about 50 percent solids, of which 60 percent or more is strongly emulsified fat, and are used in bakery foods for their effect on colour, flavour, and texture. Egg whites, containing only about 12 percent solids, primarily protein, and no fat, are important primarily for their texturizing function and give foams of low density and good stability when beaten. When present in substantial amounts, they tend to promote small, uniform cell size and relatively large volume. Meringues and angel food cakes are dependent on egg white foams for their basic structure. Although fats and oils greatly diminish its foaming power, the white still contributes to the structure of layer cakes and similar confections containing substantial amounts of both shortening and egg products.

Egg products are available to bakers in frozen or dried form. Few commercial bakers break fresh eggs for ingredients because of labour costs, unstable market conditions, and sanitary considerations. Many bakers use dried egg products because of their greater convenience and superior storage stability over frozen eggs. Processed and stored correctly, dried egg products are the functional equivalent of the fresh material, although flavour of the baked goods may be affected adversely at very high usage levels.

Sweeteners. Normal wheat flour contains about 1 percent sugars. Most are fermentable compounds, such as sucrose, maltose, glucose, and fructose. Additional maltose is formed during fermentation by the action of amyloytic enzymes (from malt and flour) on the starch. Glucose and sucrose are the sugars most frequently added to doughs and batters. The action of yeast rapidly converts the sucrose to fructose and glucose (*i.e.,* invert sugar). Invert sugar can also be added.

Effects of sugar on doughs

Sweetening power is an important property of added sugars, but sugars also provide fermentables for yeast activity. Crust colour development is related to the amount of reducing sugars present, and a dough in which the sugars have been thoroughly depleted by yeast will produce a pale crust.

Doughs with high concentrations of dissolved substances retard fermentation because of the effect on yeast of the high osmotic pressure (low water activity) of the aqueous phase. Sugars constitute the bulk of dissolved materials in most doughs. For this reason, sweet yeast-leavened goods develop gas and expand more slowly than bread doughs.

YEAST-LEAVENED PRODUCTS

Breads and rolls. Most of the bakery foods consumed throughout the world are breads and rolls made from yeast-leavened doughs. The yeast-fermentation process leads to the development of desirable flavour and texture, and such products are nutritionally superior to products of the equivalent chemically leavened doughs, since yeast cells themselves add a wide assortment of vitamins and good quality protein.

White bread. Satisfactory white bread can be made from flour, water, salt, and yeast. (A "sourdough" addition may be substituted for commercial yeast.) Yeast-raised breads based on this simple mixture include Italian-style bread and French or Vienna breads. Such breads have a hard crust, are relatively light in colour, with a coarse and tough crumb, and flavour that is excellent in the fresh bread but deteriorates in a few hours. In the United States, commercially produced breads of this type are often modified by the addition of dough improvers, yeast foods, mold inhibitors, vitamins, minerals, and small quantities of enriching materials such as milk solids or shortening. Formulas may vary greatly from one bakery to another and between different sections of the country. The standard low-density, soft-crust bread and rolls constituting the major proportion of breads and rolls sold in the United States contain greater quantities of enriching ingredients than the lean breads described above.

Whole wheat bread. Whole wheat bread, using a meal made substantially from the entire wheat kernel instead of flour, is a dense, rather tough, dark product. Breads sold as wheat or part-whole-wheat products contain a mixture of whole grain meal with sufficient white flour to produce satisfactory dough expansion.

Rye bread. Bread made from crushed or ground whole rye kernels, without any wheat flour, such as pumpernickel, is dark, tough, and coarse-textured. Rye flour with the bran removed, when mixed with wheat flour, allows production of a bread with better texture and colour. In darker bread it is customary to add caramel colour to the dough. Most rye bread is flavoured with caraway seeds.

Potato bread. Potato bread, another variety that can be leavened with a primary ferment, was formerly made with a sourdough utilizing the action of wild yeasts on a potato mash and producing the typical potato-bread flavour. It is now commonly prepared from a white bread formula to which potato flour is added.

Sweet breads. *Ingredients.* Sweet goods made from mixtures similar to bread doughs include "raised" doughnuts, Danish pastries, and coffee cakes. Richer in shortening, milk, and sugar than bread doughs, sweet doughs often contain whole eggs, egg yolks, egg whites, or corresponding dried products. The enriching ingredients alter the taste, produce flakier texture, and improve nutritional quality. Spices such as nutmeg, mace, cinnamon, coriander, and ginger are frequently used for sweet-dough products; other common adjuncts include vanilla, nuts and nut pastes, peels or oils of lemon or orange, raisins, candied fruit pieces, jams, and jellies.

Danish dough. Although various portion-size sweet goods are often called "Danish pastry," the name originally referred only to products made by a special roll-in procedure, in which yeast-leavened dough sheets are interleaved with layers of butter and the layers are reduced in thickness, then folded and resheeted to obtain many thin layers of alternating shortening and dough. Danish doughs ordinarily receive little fermentation. Before the

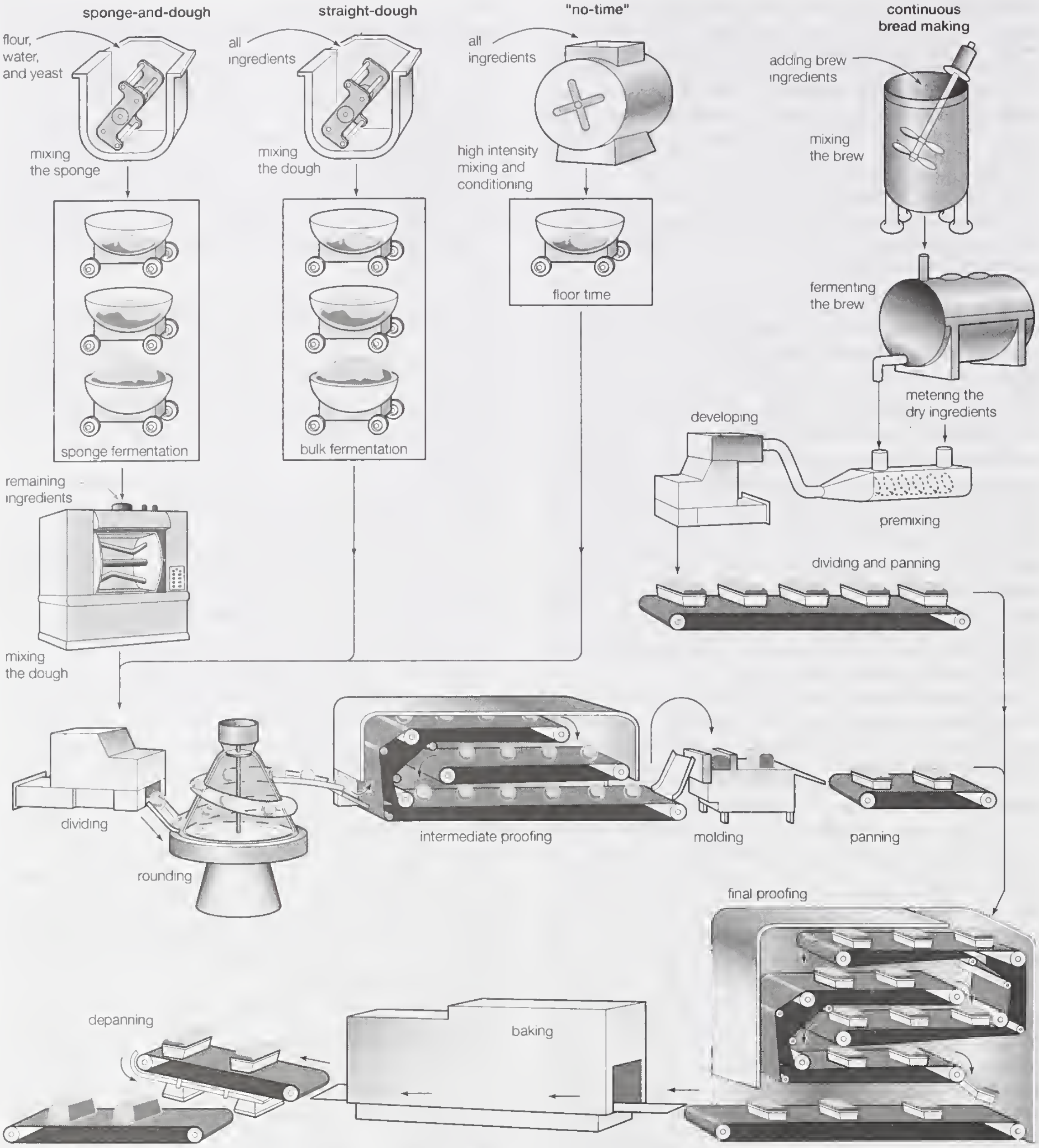


Figure 14: Essential steps in four commercial bread-making processes.
Encyclopædia Britannica, Inc.

fat is rolled in, there is a period of 20 to 30 minutes in the refrigerator, allowing gas and flavour to develop. Proof time, fermentation of the piece in its final shape, is usually only 20 to 30 minutes, at lower temperatures. When properly made, these doughs yield flaky baked products, rich in shortening, with glossy crusts.

Dough preparation. The process most commonly employed in preparing dough for white bread and many specialty breads is known as the sponge-and-dough method, in which the ingredients are mixed in two distinct stages. Another conventional dough-preparation procedure, used commonly in preparing sweet doughs but rarely regular bread doughs, is the straight-dough method, in which all the ingredients are mixed in one step before fermentation. In a less conventional method, known as the "no-time" method, the fermentation step is eliminated entirely. These processes are described below and illustrated in Figure 14.

The sponge-and-dough method. The sponge-and-dough mixing method consists of two distinct stages. In the first stage, the mixture, called the sponge, usually contains one-half to three-fourths of the flour, all of the yeast, yeast foods, and malt, and enough water to make a stiff dough. Shortening may be added at this stage, although it is usually added later, and one-half to three-fourths of the salt may be added to control fermentation. The sponge is customarily mixed in a large, horizontal dough mixer (see Figure 16), processing about one ton per batch, and usually constructed with heat-exchange jackets, allowing temperature control. The objectives of mixing are a nearly homogeneous blend of the ingredients and "developing" of the dough by formation of the gluten into elongated and interlaced fibres that will form the basic structure of the loaf. Because intense shearing actions must be avoided, the usual dough mixer has several horizontal bars, oriented parallel to the body of the mixer, rotating slowly at 35 to 75 revolutions per minute, stretching and kneading the dough by their action. A typical mixing cycle would be about 12 minutes.

Dough mixers

The mixed sponge is dumped into a trough (Figure 15), a shallow rectangular metal tank on wheels, and placed in an area of controlled temperature and humidity (*e.g.*, 27° C [80° F] and 75 percent relative humidity), where it is fermented until it begins to decline in volume. The time required for this process, called the drop or break, depends on such variables as temperature, type of flour, amount of yeast, absorption, and amount of malt, which

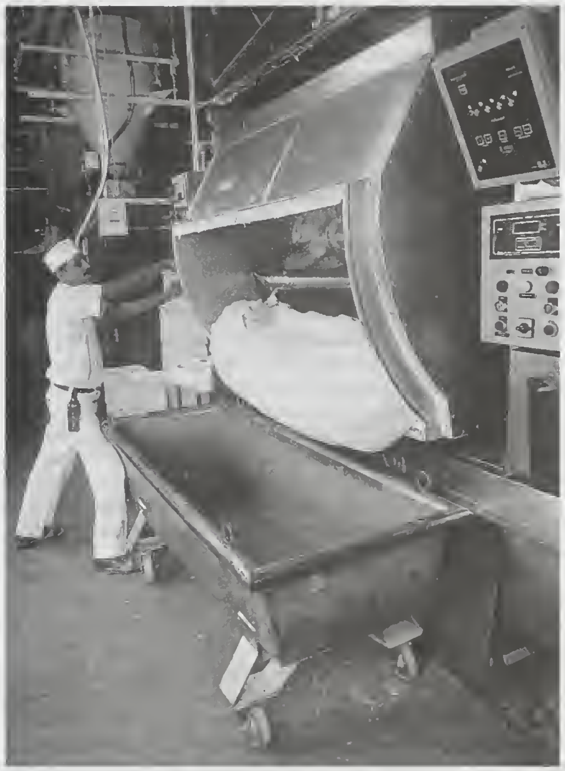


Figure 15: Transferral of bread "sponge" from dough mixer to trough prior to fermentation.
© Mathew Neal McVay/Tony Stone Images

are frequently adjusted to produce a drop in about three to five hours.

At the second, or dough, stage, the sponge is returned to the mixer, and the remaining ingredients are added. The dough is developed to an optimum consistency, then either returned to the fermentation room or allowed "floor time" for further maturation.

Advantages of the sponge-and-dough method include: (1) a saving in the amount of yeast (about 20 percent less is required than for a straight dough), (2) greater volume and more desirable texture and grain, and (3) greater flexibility allowed in operations because, in contrast to straight doughs (which must be taken up when ready), sponges can be held for later processing without marked deterioration of the final product.

The sponge method, however, involves extra handling of the dough, additional weighing and measuring, and a second mixing and thus has the disadvantage of increasing labour, equipment, and power costs.

The straight-dough method. Two of the many possible variations in the straight-dough process include the remixed straight-dough process, with a small portion of the water added at the second mix, and the no-punch method, involving extremely vigorous mixing. The straight-dough method is rarely used for white breads because it is not sufficiently adaptable to allow compensation for fluctuations in ingredient properties.

"No-time" methods. One set of procedures intended to eliminate the traditional bulk fermentation step are the "no-time" methods. Popular in the United Kingdom and Australia, these processes generally require an extremely energy-intensive mixing step, sometimes performed in a partially vacuumized chamber. Rather high additions of chemical oxidants, reducing agents, and other dough modifiers are almost always required in order to produce the desired physical properties. "No-time" is actually a misnomer, since there are always small amounts of floor time (periods when the dough is awaiting further processing) during which maturing actions lead to improvements in the dough's physical properties. Even then, the flavour of the bread cannot be expected to match that of a traditionally processed loaf.

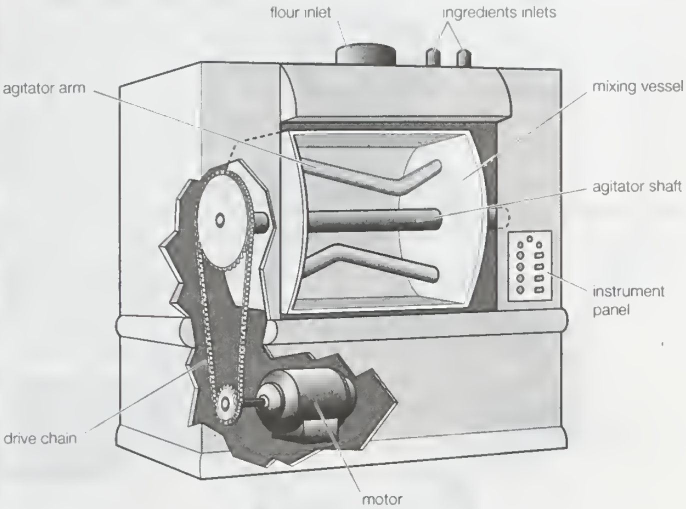


Figure 16: The drive and mixing elements of a horizontal mixer.
Encyclopædia Britannica, Inc.

Makeup. After the mass of dough has completed fermentation (and has been remixed if the sponge-and-dough process is employed), it is processed by a series of devices loosely classified as makeup equipment. In the manufacture of pan bread, makeup equipment includes the divider, the rounder, the intermediate proofer, the molder, and the panner.

Dividing. The filled trough containing remixed dough is moved to the divider area or to the floor above the divider. The dough is dropped into the divider hopper, which cuts it into loaf-size pieces (see Figure 17). Two methods are employed. In the volumetric method, the dough is forced into pockets of a known volume. The pocket contents are cut off from the main dough mass and then ejected onto a conveyor leading to the rounder. When density is kept constant, weight and volume of the dough pieces are roughly the same. In the weight-based method, a cylindrical rope of dough is continuously extruded through an orifice at a fixed rate and is cut off by a knife-edged rotor at fixed intervals. Since the dough is of consistent density, the cut pieces are of uniform weight. Like the pocket-cut pieces, the cylindrical pieces are conveyed to the rounder.

Rounding. Dough pieces leaving the divider are irregular in shape, with sticky cut surfaces from which the gas can readily diffuse. Their gluten structure is somewhat disoriented and unsuitable for molding. The rounder closes these cut surfaces, giving each dough piece a smooth and

Function of the rounder

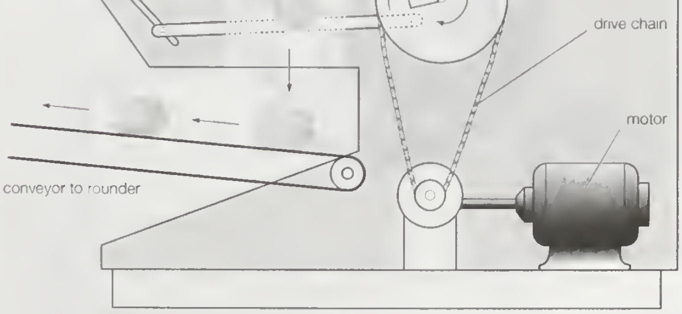


Figure 17: The drive, feed, and cutting elements of a dough divider.
Encyclopædia Britannica, Inc

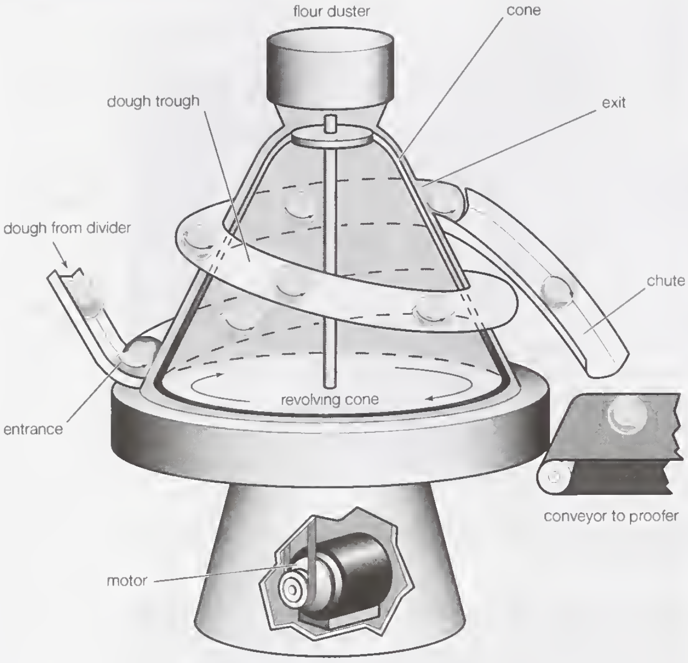


Figure 18: The movement of dough through a dough rounder.
Encyclopædia Britannica, Inc.

dry exterior; forms a relatively thick and continuous skin around the dough piece, reorienting the gluten structure; and shapes the dough into a ball for easier handling in subsequent steps. It performs these functions by rolling the well-floured dough piece around the surface of a drum or cone, moving it upward or downward along this surface by means of a spiral track (see Figure 18). As a result of this action, the surface is dried both by the even distribution of dusting flour and by dehydration resulting from exposure to air; the gas cells near the surface of the ball are collapsed, forming a thick layer inhibiting the diffusion of gases from the dough; and the dough piece assumes an approximately spherical shape.

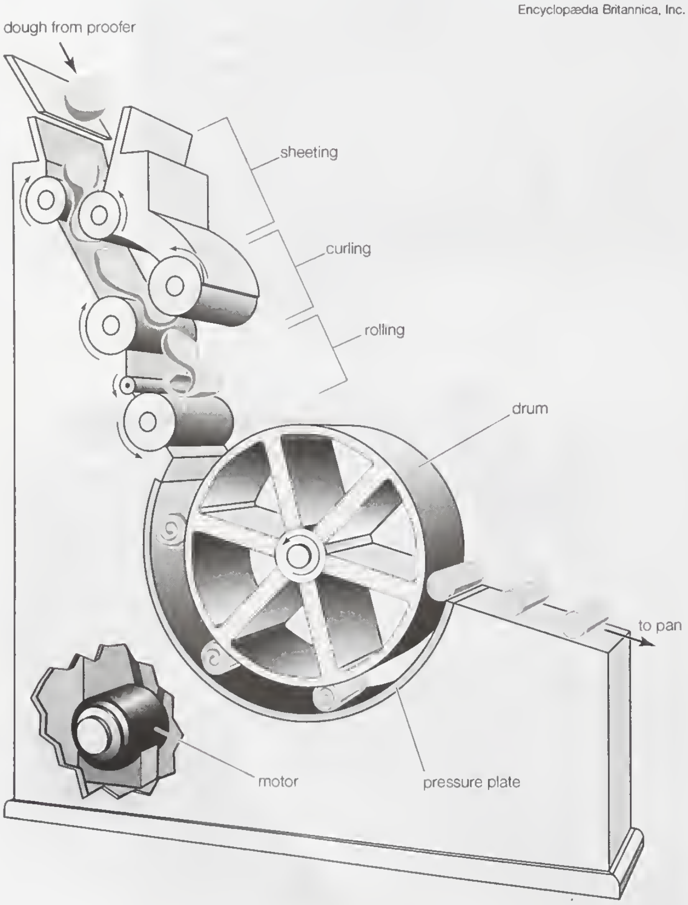


Figure 19: The formation of dough cylinders in a drum molder.
Encyclopædia Britannica, Inc.

Intermediate proofing. Dough leaving the rounder is almost completely degassed. It lacks extensibility, tears easily, has rubbery consistency, and has poor molding properties. To restore a flexible, pliable structure, the dough piece must be allowed to rest while fermentation proceeds. This is accomplished by letting the dough ball travel through an enclosed cabinet, the intermediate proofer, for several minutes. Physical changes, other than gas accumulation, occurring during this period are not yet understood, but there are apparently alterations in the molecular structure of the dough rendering it more responsive to subsequent operations. Upon leaving the intermediate proofer, the dough is more pliable and elastic, its volume is increased by gas accumulation, and its skin is firmer and drier.

Restoring dough structure

Most intermediate proofers are the overhead type, in which the principal part of the cabinet is raised above the floor, allowing space for other makeup machinery beneath it. Interior humidity and temperature depend on humidity accumulating from the loaves and on ambient temperatures.

Molding. The molder receives pieces of dough from the intermediate proofer and shapes them into cylinders ready to be placed in the pans. There are several types of molders, but all have four functions in common: sheeting, curling, rolling, and sealing (see Figure 19). The dough as it comes from the intermediate proofer is a flattened spheroid; the first function of the molder is to flatten it into a thick sheet, usually by means of two or more consecutive pairs of rollers, each succeeding pair set more closely together than the preceding pair. The sheeted dough is curled into a loose cylinder by a special set of rolls or by a pair of canvas belts. The spiral of dough in the cylinder is not adherent upon leaving the curling section, and the next operation of the molder is to seal the dough piece, allowing it to expand without separating into layers. The conventional molder rolls the dough cylinder between a large drum and a smooth-surfaced semicircular compression board. Clearance between the drum and board is gradually reduced, and the dough, constantly in contact with both surfaces, becomes transversely compressed.

Panning. An automatic panning device is an integral part of most modern molders. As empty pans, carried on a conveyor, pass the end of the machine, the loaves are transferred from the molder and positioned in the pans by a compressed air-operated device. Before the filled pans are taken to the oven, the dough undergoes another fermentation, or pan-proofing, for about 20 minutes at temperatures of 40° to 50° C, or 100° to 120° F.

Continuous bread making. Many steps in conventional dough preparation and makeup have been fully automated, but none of the processes is truly continuous. In continuous systems, the dough is handled without interruption from the time the ingredients are mixed until it is deposited in the pan. The initial fermentation process is still essentially a batch procedure, but in the continuous bread-making line the traditional sponge is replaced by a liquid pre-ferment, called the broth or brew. The brew consists of a mixture of water, yeast, sugar, and portions of the flour and other ingredients, fermented for a few hours before being mixed into the dough.

Fermenting a liquid brew

As shown in Figure 14, after the brew has finished fermenting, it is fed along with the dry ingredients into a mixing device, which mixes all ingredients into a homogeneous mass. The batterlike material passes through a dough pump regulating the flow and delivering the mixture to a developing apparatus, where kneading work is applied. The developer is the key equipment in the continuous line. Processing about 50 kilograms (100 pounds) each 90 seconds, it changes the batter from a fluid mass having no organized structure, little extensibility, and inadequate gas retention to a smooth, elastic, film-forming dough. The dough then moves out of the developer into a metering device that constantly extrudes the dough and intermittently severs a loaf-size piece, which falls into a pan passing beneath.

Although ingredients are generally the same as those used in batch processes, closer control and more rigid specifications are necessary in continuous processing in order to assure the satisfactory operation of each unit. Changes in conditions cannot readily be made to compensate for changes occurring in ingredient properties. Oxidizers, such as bromate and iodate, are added routinely to compensate for the smaller amount of oxygen brought into the dough during mixing.

The use of fermented brews has been widely accepted in plants practicing traditional dough preparation and makeup. The handling of a fermentation mixture through pumps, pipes, valves, and tanks greatly increases efficiency and control in both batch-type and continuous systems.

Baking and depanning. *Ovens.* The output of all bread-making systems, batch or continuous, is usually keyed to the oven, probably the most critical equipment in the bakery. Most modern commercial bakeries use either the tunnel oven, consisting of a metal belt passing through a connected series of baking chambers open only at the ends, or the tray oven, with a rigid baking platform carried on chain belts. Other types include the peel oven, having a fixed hearth of stone or brick on which the loaves are placed with a wooden paddle or peel; the reel oven, with shelves rotating on a central axle in Ferris wheel fashion; the rotating hearth oven; and the draw plate oven.

Advances in high-capacity baking equipment include a chamber oven with a conveyor that carries pan assemblies (called straps) along a roughly spiral path through an insulated baking chamber. The straps are automatically added to the conveyor before it enters the oven and then automatically removed and the bread dumped at the conveyor's exit point. Although the conveyor is of a complex design, the oven as a whole is considerably simpler than most other high-capacity baking equipment and can be operated with very little labour. As a further increase in efficiency, the conveyor can also be designed to carry filled pans in a continuous path through a pan-proofing enclosure and then through the oven.

The rack oven for small bakeries

In small to medium-size retail bakeries, baking may be done in a rack oven. This consists of a chamber, perhaps two to three metres high, that is heated by electric elements or gas burners. The rack consists of a steel framework having casters at the bottom and supporting a vertical array of shelves. Bread pans containing unbaked dough pieces are placed on the shelves before the rack is pushed mechanically or manually into the oven. While baking is taking place, the rack may remain stationary or be slowly rotated.

Most ovens are heated by gas burned within the chamber, although oil or electricity may be used. Burners are sometimes isolated from the main chamber, heat transfer then occurring through induced currents of air. Baking reactions in the oven are both physical and chemical in nature. Physical reactions include film formation, gas expansion, reduction of gas solubility, and alcohol evaporation. Chemical reactions include yeast fermentation, carbon dioxide formation, starch gelatinization, gluten coagulation, sugar caramelization, and browning.

Depanners. Automatic depanners, removing the loaves from the pans, either invert the pans, jarring them to dislodge the bread, or pick the loaves out of the pans by means of suction cups attached to belts.

CHEMICALLY LEAVENED PRODUCTS

Many bakery products depend on the evolution of gas from added chemical reactants as their leavening source. Items produced by this system include layer cakes, cookies, muffins, biscuits, corn bread, and some doughnuts.

The gluten proteins of the flour serve as the basic structural element in chemically leavened foods, just as they do in bread. The relatively smaller amounts of flour, the weaker (less extensible) protein in the soft-wheat flours customarily employed, and the lower protein content of the flour, however, result in a softer, crumblier texture. In most chemically leavened foods, the protein content of the flour, inadequate in quantity and quality to support the amount of expansion required in bread, produces a product of higher density.

Prepared mixes and doughs. Prepared dry mixes, available for home use and for small and medium-size commercial bakeries, vary in complexity from self-rising flour, consisting only of salt, leavening ingredients, and flour, to elaborate cake mixes. Mixes offer the consumer ingredients measured with greater accuracy than possible with kitchen utensils and special ingredients designed for functional compatibility.

Prepared doughs for such products as biscuits and other quick breads, packaged in cans of fibre and foil laminates, are available in refrigerated form. These products carry the mix concept two steps further; the dough or batter is premixed and shaped. Unlike ordinary canned products, refrigerated doughs are not sterile but contain microbes from normal ingredient contamination. Spoilage is retarded by low storage temperature, low oxygen tension, and the high osmotic pressure of the aqueous phase.

Frozen batters

Many boutique cookie bakeries and muffin shops that operate in shopping malls and similar locations generally use frozen batters shipped from a central plant. These batters are thawed a day or so before use, and a measured amount is scooped from the container and placed on a baking pan immediately before insertion into the oven. In this way, freshly baked cookies or muffins can be prepared in many varieties with a small amount of unskilled labour and a minimum of specialized equipment. In some cases, a central commissary supplies fully baked but frozen products, which are simply thawed (and sometimes iced and decorated) before sale.

Dough and batter formulas. *Hot breads.* Hot breads, such as biscuits, muffins, pancakes, and scones, constitute a large and important class of chemically leavened bakery foods. They consist of flour, baking powder, salt, and liquid, with varying amounts of eggs, milk, sugar, and shortening. Other variations include the addition of fruits such as raisins, condiments such as peppers, and adjuncts such as cheese. In corn breads a considerable proportion of the flour is replaced by cornmeal. Mixing and forming methods, and the baking conditions applied, also affect product appearance, texture, and flavour. For example, a batter suitable for making corn bread might also be used to make muffins or pancakes, and each kind of finished product would vary not only in appearance but also in flavour and texture. Recipes for hot breads usually contain not more than about 15 percent shortening and 5 percent sugar. Eggs, when used, are customarily whole eggs. Milk is often used both for flavour and for its texturizing and crust coloration properties.

Cakes. There are traditional rules for assuring "formula balance," or the correct proportioning of ingredients, in layer cakes. For every 10 parts of flour, yellow layer cakes should contain 10 to 16 parts sugar by weight, and white layer cakes should contain 11 to 16 parts sugar. Shortening should range from 3 to 7 parts for each 10 parts of flour. The weight of liquid whole eggs should equal or exceed that of the shortening in the mixture. Total water, including the moisture in eggs and milk, should exceed the amount of sugar by $2^{1}/_{2}$ to $3^{1}/_{2}$ parts. Baking-powder weight should equal from 3 to 6 percent of flour weight; salt should equal 3 to 4 percent of flour weight. If the amount of sugar in a formula is increased, the egg content should be increased an equal amount, and more shortening should be added when the percentage of eggs is increased. Additional water is rarely added when the formula contains dry milk, but if the formula water is not sufficient to equal the reconstitution water for the milk, about 1 percent of water for each additional percent of milk solids is added.

Common cake varieties include white cake, similar in formula to yellow cake, except that the white cake uses egg whites instead of whole eggs; devil's food cake, differing from chocolate cake chiefly in that the devil's food batter is adjusted to an alkaline level with sodium bicarbonate; chiffon cakes, deriving their unique texture from the effect of liquid shortening on the foam structure; and gingerbread, similar to yellow cake but containing large amounts of molasses and spices.

Cookies. Recipes for cookies (called biscuits or sweet biscuits in some countries) are probably more variable than those for any other type of bakery product. Some layer-cake batters can be used for soft drop cookies, but most cookie formulas contain considerably less water than cake recipes, and cookies are baked to a lower moisture content than any normal cake. With the exception of soft types, the moisture content of cookies will be below 5 percent after baking, resulting in crisp texture and good storage stability.

Low moisture content

Cookies are generally high in shortening and sugar. Milk and eggs are not common ingredients in commercial cookies but may be used in home recipes. Sugar granule size has a pronounced effect on cookie texture, influencing spread and expansion during baking, an effect partly caused by competition for the limited water content between the slowly dissolving sugar and the gluten of the flour.

Equipment. *Mixing.* The horizontal dough mixers used for yeast-leavened products may be used for mixing chemically leavened doughs and batters. Mixers may be the batch type, similar in configuration to the household mixer, with large steel bowls, open at the top, containing the batter while it is mixed or whipped by beater paddles of various conformations. In continuous mixers the batter is pumped through an enclosed chamber while a toothed disk rapidly rotates and mixes the ingredients. The chambers may be pressurized to force gas into the batter and surrounded with a flowing heat-transfer medium to adjust the temperature.

Sheeting and cutting. Chemically leavened doughs can be formed by methods similar to those used for yeast-leavened doughs of similar consistency. In the usual sequence, the dough passes between sets of rollers, forming sheets of uniform thickness; the desired outline is cut in the sheet by stamping pressure or embossed rollers; and the scrap dough is removed for reprocessing. Many cookies and crackers are made in this way, and designs may be impressed in the dough pieces by docking pins (used primarily to puncture the sheet, preventing formation of excessively large gas bubbles) or by cutting edges partially penetrating the dough pieces.

Die forming and extruding. In addition to the sheeting and cutting methods, cookies may be shaped by die forming and extrusion. In die forming a dough casing may be applied around a centre portion of jam or other material, forming products such as fig bars; or portions of dough may be deposited, forming such drop-type cookies as vanilla wafers, chocolate chip, and oatmeal cookies. Extrusion is accomplished by means of a die plate having orifices that may be circular, rectangular, or complex in outline. The mass of dough, contained in a hopper, is pushed through these openings, forming long strands of dough. Individual cookies are formed by separating pieces from the dough strand with a wire passing across the outer surface of the die or by pulling apart the hopper and oven belt (to which the dough adheres).

Extrusion of cookies

Rotary molding. Cookies produced on rotary molders include sandwich-base cakes and pieces made with embossed designs. A steel cylinder, the surface covered with shallow engraved cavities, rotates past the opening in a hopper filled with cookie dough. The pockets are filled with the dough, which is sheared off from the main mass by a blade, and, as the cylinder continues its revolution, the dough pieces are ejected onto a conveyor belt leading to the band oven.

Baking. Most commercial ovens for chemically leavened products are the band types, although reel ovens are still used, especially in smaller shops or bakeries where short runs are frequent.

AIR- AND STEAM-LEAVENED PRODUCTS

Air leavening. Air-leavened bakery products, avoiding the flavours arising from chemical- and yeast-leavening systems, are particularly suitable for delicately flavoured cakes. Since the batters can be kept on the acidic side of neutrality, the negative influence of chemical leaveners on fruit flavours and vanilla is avoided.

Foaming action of protein

Foams and sponges. The albumen of egg white, a protein solution, foams readily when whipped. The highly extended structure has little strength and must be supported during baking by some other protein substance, usually the gluten of flour. Because the small amount of lipids in flour tend to collapse the albumen foam, flour is gently folded into egg white foams, minimizing contact of fatty substances with the protein. Gluten sponges are denser than the lightest egg-white foams but are less subject to fat collapse.

The foam of egg yolks and whole eggs, as in pound cakes, is an air-in-oil emulsion. Proteins and starch, scattered throughout the emulsion in a dispersed condition, gradually coalesce as the batter stands or is heated. Fats and oils, in addition to yolk lipids, can be added to such systems without causing complete collapse but never achieve the low density possible with protein foams and usually have a tender, crumbly texture, unlike the more elastic structure of albumen-based products.

Wafers and biscuits. Rye wafers made of whipped batters are modern versions of an ancient Scandinavian food. High-moisture dough or batter, containing a substantial amount of rye flour and some wheat flour, is whipped, extruded onto an oven belt, scored and docked, then baked slowly until almost dry. Alternatively, the strips of dough may be cut after they are baked.

Beaten biscuits, an old specialty of the American South, are also made from whipped batter. Air is beaten into a stiff folded dough with many strokes of a rolling pin or similar utensil. Round pieces cut from the dough are pricked with a fork to prevent development of large bubbles, then baked slowly. The baked biscuit is similar to a soft cracker.

Steam leavening. All leavened products rely to some extent on water-vapour pressure to expand the vesicles or gas bubbles during the latter stages of baking, but some items also utilize the leavening action produced by the rapid buildup of steam as the interior of the product reaches the boiling point. These foods include puff pastries, used for patty shells and napoleons, and chou pastes, often used for cream-puff and éclair cases.

Puff pastry. Puff pastry, often used in French pastries, is formed from layered fat and dough. The proportion of fat is usually high, rarely less than 30 percent of the finished raw piece. The dough should be extensible but not particularly elastic; for this reason mixtures of hard and soft wheat flour are often used. The fat should have an almost waxy texture and must remain solid through the sheeting steps. Butter, although frequently used, is not particularly suitable for puff pastry because its low melting point causes it to blend into the dough during the sheeting process. Bakers specializing in puff pastry often use special margarines containing high-melting-point fats.

Puff pastry production methods

There are several methods of making puff pastry. In the basic procedure dough is rolled into a rectangular layer of uniform thickness, and the fat is spread over two-thirds of the surface. The dough is next folded, producing three dough strata enclosing two fat layers. This preparation is next chilled by refrigeration, then rolled, reducing thickness until it reaches approximately the area of the original unfolded dough. The folding, refrigeration, and rolling procedure is repeated several times, and after the final rolling the dough is reduced to the thickness desired in the shaped raw piece.

Correctly prepared puff pastry will expand as much as 10 times during baking because of the evolution of large volumes of steam at the interface between shortening and dough. The focuses for gassing are the microscopic air bubbles rolled into the dough during the layering process. If layering has been properly conducted, the finished pieces will be symmetrical and well-shaped, with crisp, flaky outer layers.

Chou paste. Chou paste, used for cream puffs, is made by an entirely different method. Flour, salt, butter, and boiling water are mixed together, forming a fairly stiff dough, and whole eggs are incorporated by beating. Small pieces of the dough are baked on sheets, initially at high temperature. The air bubbles formed during mixing expand rapidly at baking temperatures, filling the interior with large, irregular cells, while the outside browns and congeals, forming a rather firm case. The interior, largely hollow, can be injected with such sweet or savoury fillings as whipped cream or shrimp in sauce.

MARKET PREPARATION

Slicing. Bread often is marketed in sliced form. Slicing is performed by parallel arrays of saw blades through which the loaves are carried by gravity or by conveyors. The blades may be endless bands carried on rotating drums, or relatively short strips held in a reciprocating frame. Most bread is sliced while still fairly warm, and the difficulty of cutting the sticky, soft crumb has led to development of coated blades and blade-cleaning devices. Horizontal slicing of hamburger rolls and similar products is accomplished by circular (disk) blades, usually two blades in a slicer, between which a connected array of four or six rolls is carried by a belt. The cutter blades are separated to avoid cutting completely through the roll, in order to leave a "hinge."

Quality control in freezing

Freezing. Freezing is an indispensable bakery industry process. Ordinary bread and rolls are rarely distributed and sold in frozen form because of the excessive cost in relation to product value, but a substantial percentage of all specialty products is sold in frozen form. Most bakery products respond well to freezing, although some cream fillings must be specially formulated to avoid syneresis, or gel breakdown. Rapid chilling in blast freezers is preferred, although milder methods may be used. Storage at −18° C (0° F) or lower is essential for quality maintenance. Thawing and refreezing is harmful to quality. Frozen bakery products can dehydrate under freezer conditions and must be packaged in containers resistant to moisture-vapour transfer.

Wrapping. Most American consumers prefer wrapped bread, and the trend toward wrapping is growing in other countries. Sanitary and aesthetic considerations dictate protection of the product from environmental contamination during distribution and display. Waxed paper was originally the only film used to package bread; then cellophane became popular; and finally polyethylene, polypropylene, and combination laminates became common. Other bakery products are packaged in a variety of containers ranging from open bags of greaseproof material to plastic trays with sealed foil overwraps.

QUALITY MAINTENANCE

Spoilage by microbes. Bakery products are subject to the microbiological spoilage problems affecting other foods. If moisture content is kept below 12 to 14 percent (depending on the composition), growth of yeast, bacteria, and molds is completely inhibited. Nearly all crackers and cookies fall below this level, although jams, marshmallow, and other adjuncts may be far higher in moisture content. Breads, cakes, sweet rolls, and some other bakery foods may contain as much as 38 to 40 percent water when freshly baked and are subject to attack by many fungi and a few bacteria.

Fungi. To obtain maximum shelf life free of mold spoilage, high levels of sanitation must be maintained in baking and packing areas. Oven heat destroys all fungal life-forms, and any spoilage by these organisms is due to reinoculation after baking. A number of compounds have been proposed for use as fungistats in bread. Some have proved to be innocuous to molds, toxic to humans, or both. Soluble salts of propionic acid, principally sodium propionate, have been accepted and extend shelf life appreciably in the absence of a massive inoculum. Sorbic acid (or potassium sorbate) and acetic acid also have a protective effect.

The only widespread food poisoning in which bread has been a vector has resulted from ergot, a fungus infection of the rye plant. Ergot contamination of bread made from rye, or from blends of rye and wheat, has caused epidemics leading to numerous deaths.

Bacteria. Bacteria associated with bread spoilage include *Bacillus mesentericus*, responsible for "ropy" bread, and the less common but more spectacular *Micrococcus prodigiosus*, causative agent of "bleeding bread." Neither ropy bread nor bleeding bread is particularly toxic. Enzymes secreted by *B. mesentericus* change the starch inside the loaf into a gummy substance stretching into strands when a piece of the bread is pulled apart. In addition to ropiness, the spoiled bread will have an off-aroma sometimes characterized as fruity or pineapple-like. Formerly, when ropiness occurred, bakers acidified doughs with vinegar as a protective measure, but this type of spoilage is rare in bread from modern bakeries.

M. prodigiosus causes red spots to appear in bread. At an advanced stage these spots of high bacterial population may liquefy, emphasizing the similarity to blood, which has sometimes led the superstitious to attribute religious significance to this phenomenon. The organism will not survive ordinary baking temperatures, unlike *B. mesentericus*, which forms spores capable of survival in the centre of the loaf, where the temperature rises only to about 100° C (212° F).

Baked goods containing such high-moisture adjuncts as pastry creams and pie fillings are susceptible to contamination by food-spoilage organisms, including *Salmonella* and *Streptococcus*. Cream and custard pies are recognized health hazards when stored at room temperature for any length of time, and some communities ban their sale during summer. Storage in frozen form eliminates the hazard.

Structural changes in staling

Staling. Undesirable changes in bakery products can occur independently of microbial action. Staling involves changes in texture, flavour, and appearance. Firming of the interior, or "crumb," is a highly noticeable alteration in bread and other low-density, lean products. Elasticity is lost, and the structure becomes crumbly. Although loss of moisture produces much the same effect, texture staling can occur without any appreciable drying. Such firming is due to changes in the molecular status of the starch, specifically to a kind of aggregation of sections of the long-chain molecules into micelles, making the molecules more rigid and less soluble than in the newly gelatinized granule. Bread that has undergone texture staling can be softened by heating to about 60°–65° C (140°–150° F). However, its texture does not return to that of fresh bread, being gummier and more elastic. In addition, care must be exercised to prevent drying during heating.

Starch retrogradation, the cause of ordinary texture staling of the crumb, can be slowed by the addition of certain compounds to the dough. Most of the effective chemicals are starch-complexing agents. Monoglycerides of fatty acids have been widely used as dough additives to retard staling in the finished loaf. (S.A.M.)

Fruits

Fruit is sometimes defined as the product of growth from an angiosperm, or flowering plant. From a purely botanical point of view, the fruit may be only the fleshy growth that arises from the ovary of a flower and may not necessarily include any other structures. From the consumer's or food processor's point of view, however, fruit is generally characterized as the edible product of a plant or tree that includes the seed and its envelope and can typically be described as juicy, sweet, and pulpy. Typical fruit structures are illustrated in Figure 20.

Fruits are a high-moisture, generally acidic food that is relatively easy to process and that offers a variety of flavour, aroma, colour, and texture to the diet. They are usually low in calories but are an excellent source of dietary fibre and essential vitamins. Owing to the presence of cellulose, pectin, and various organic acids, fruits can also act as natural laxatives. Fruits are therefore a valuable part of the diet.

FRUIT CHARACTERISTICS

Nutrient composition. *Moisture content, acidity, and vitamin content.* As shown in Table 4, fresh fruit is typically between 75 and 95 percent water, a fact that helps to explain the refreshing character of the food. In general, fruits are acidic, with pH ranging from 2.5 to 4.5. The most common acids in fruits are citric acid, malic acid, and tartaric acid.

Vitamin C

Of all the vitamins present in fruits, the most noted is vitamin C, or ascorbic acid. Actual quantities of vitamin C in fruits are not especially large, but the vitamin is particularly important in the diet because of its role in the prevention of disease and in the general promotion of good health. Citrus fruits, such as oranges, lemons, and

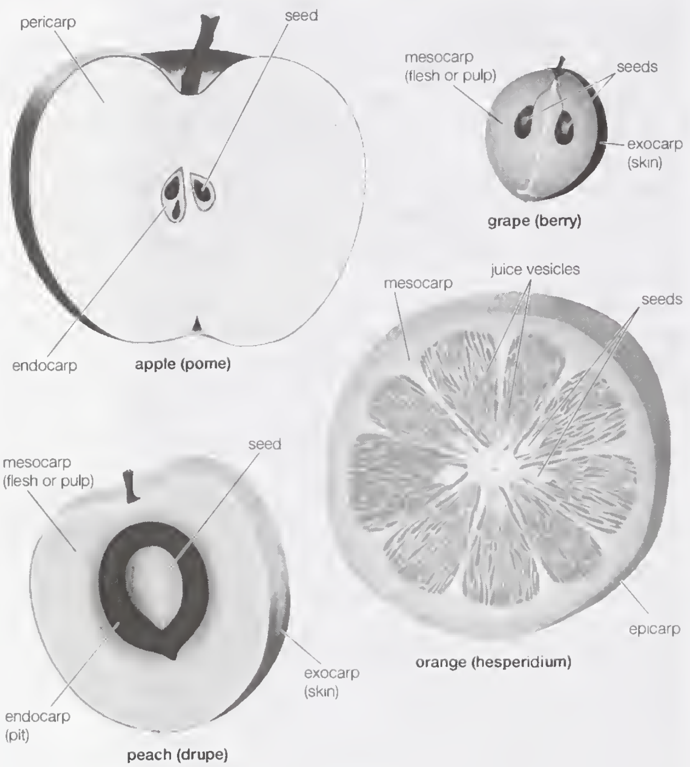


Figure 20: Four representative types of fruit.
Encyclopædia Britannica, Inc

grapefruits, are well known for their vitamin C content. Other sources include most berries and melons. Carotene, a chemical common to fruit, is easily converted in the body to vitamin A; cantaloupes, peaches, and apricots are significant sources of this nutrient.

Carbohydrates. Typically, fruits are high in carbohydrates, although a large range is possible—between 2 and 40 percent, depending on the type of fruit and its maturity. Free sugars usually include fructose, glucose, and sucrose; other sugars may be present in smaller quantities.

A large portion of the carbohydrates present in fruits is fibre, which is not digested and passes through the digestive system. Fibre is usually made up of cellulose, hemicellulose, and pectic substances. A small amount of starch may also be present in fruit, but starches are typically converted to sugars during the ripening process.

Protein and fat content. A negligible quantity of protein is found in fruits, and they usually contain less than 1 percent fat. Fats are most typically associated with the waxy cuticle surface of the fruit skin. Exceptions to this rule are avocados and olives, the flesh of which may contain as much as 20 percent oil.

Maturation and spoilage. *Ripening and senescence.* Fruits are living biological entities that perform a number of metabolic functions. Two functions of particular importance in fruit processing are respiration (the breaking down of carbohydrates, giving off carbon dioxide and heat) and transpiration (the giving off of moisture). Once the fruit is harvested, respiration and transpiration continue, but only for as long as the fruit can draw on its own food reserves and moisture. It is this limited ability to continue vital metabolic functions that defines fruit as perishable.

Fruit development can generally be divided into three major stages: growth, maturation, and senescence. The period of growth generally involves cell division and enlargement, which accounts for the increasing size of the fruit. Maturation is usually reached just prior to the end of growth and may include flavour development and increase in sugar content (detectable as increasing sweetness). Senescence is the period when chemical synthesizing pathways give way to degradative processes, leading to aging and death of tissue. Fruit ripening is thus the result of many complex changes, some interactive but many independent of one another.

Storage concerns. As harvested fruit ages, it is particularly important to manage the temperatures under which it is stored. For example, respiration largely involves enzymatic processes, which are significantly controlled by ambient temperature. The rate of chemical change in fruit generally doubles for every increase of 10° C (at room temperature, roughly 20° F).

Changes that take place during storage as fruit begins to overripen may include extreme colour formation, development of strong off-flavours with intense aroma, softening of the flesh, onset of physiological disorders, and manifestations of disease. In addition, fruit can be injured by overcooling. Chilling injury may be manifested by pitting and browning of the surface and by pitting and darkening of the flesh.

Microorganisms can also cause problems during senescence and storage. Many bacteria and fungi, for instance, are involved in decay after harvest. Typical fungi include *Alternaria, Botrytis, Monilinia, Penicillium,* and *Rhizopus.* These fungi are generally weak pathogens, in that they usually invest only weak or damaged fruit. Efforts to control infection begin in the orchard, usually with the application of fungicides. Cooling of the fruit or, conversely, hot-water dipping may also enhance storage quality. In addition, the careful application of ionizing radiation has been shown to inhibit microbial growth.

FRESH FRUIT

Storage. Once harvested, fruits are moved to storage. In the case of highly heat-sensitive products such as raspberries or cherries, the fruit should be precooled prior to storage. Precooling can be accomplished by hydrocooling (immersion of the fruit in cold water) or vacuum cooling (moistening and then placing under vacuum in order to induce evaporative cooling).

Controlled-atmosphere storage

A typical storage system for fruit is cold storage, using refrigerated air. Other techniques include controlled-atmosphere (CA) storage and hypobaric storage. In CA storage the oxygen and carbon dioxide content of the storage environment are controlled in such a way as to retard senescence and further deterioration of the fruit. In gen-

Table 4: Nutrient Composition of Selected Fruits and Fruit Products (per 100 grams)*

fruit or fruit product	energy (kcal)	water (g)	carbohydrate (g)	vitamin C (mg)	thiamine (mg)	riboflavin (mg)	niacin (mg)	vitamin A (IU)	fat (g)	protein (g)
Apple, whole	59	83.90	15.25	5.7	0.017	0.014	0.077	53	0.36	0.19
Apple juice	47	87.93	11.68	0.9	0.021	0.017	0.100	1	0.11	0.06
Apricot	48	86.35	11.12	10.0	0.030	0.040	0.600	2,612	0.39	1.40
Avocado	161	74.27	2.11	7.9	0.108	0.122	1.921	61	15.32	1.98
Banana	92	74.26	23.43	9.1	0.045	0.100	0.540	81	0.48	1.03
Grape	63	81.30	17.15	4.0	0.092	0.057	0.300	100	0.35	0.63
Grapefruit	32	90.89	8.08	34.4	0.036	0.020	0.250	124	0.10	0.63
Orange	47	86.75	11.75	53.2	0.087	0.040	0.282	205	0.12	0.94
Peach	43	87.66	11.10	6.6	0.017	0.041	0.990	535	0.09	0.70
Pear	59	83.81	15.11	4.0	0.020	0.040	0.100	20	0.40	0.39
Plum	55	85.20	13.01	9.5	0.043	0.096	0.500	323	0.62	0.79
Watermelon	32	91.51	7.18	9.6	0.080	0.020	0.200	366	0.43	0.62

*Values shown are approximations; actual nutrient composition can vary greatly depending on such factors as growing conditions, time of harvest, and storage.
Source: *Composition of Foods,* Agriculture Handbook no. 8-9, U.S. Department of Agriculture.

eral, oxygen levels are reduced and carbon dioxide levels increased. CA conditions can be generated in a number of ways. Conventional CA depends on the respiration of the fruit to generate carbon dioxide, and the concentration of this gas is controlled by wet scrubbers, hydrated lime, or other commercial carbon dioxide removal systems. Liquid nitrogen and compressed nitrogen gas have also been used to flush out the ambient air of the storage facility. In other systems oxygen is converted to carbon dioxide by reaction with liquid propane or by catalytic burning.

Hypobaric storage involves the cold storage of fruit under partial vacuum. Typical conditions include pressures as low as 80 and 40 millimetres of mercury and temperatures of 5° C (40° F). Hypobaric conditions reduce ethylene production and respiration rates; the result is an extraordinarily high-quality fruit even after months of storage.

Packaging. Packaging systems for fresh fruit usually involve a simple plastic breathable bag or overwrap. However, as the market value of high-quality fruit has increased, so too have efforts to develop improved packaging. These efforts have been primarily in the area of modified-atmosphere packaging (MAP). In this type of packaging the barrier properties of the material are carefully selected according to the respiration characteristics of the fruit. The goal is to allow an exchange of gases and moisture that produces the optimal storage environment. Continued work in this field is producing "smart" films, which not only produce the optimal atmosphere for storage but also change their barrier properties depending on the ambient temperature and on the respiration rate of the fruit.

FRUIT JUICE

After fresh fruit, one of the most common fruit products is fruit juice. Fruit juice can take on many forms, including a natural-style cloudy product, a "nectar"-type product containing suspended solids, a fully clarified juice, juice concentrate, and fruit drinks.

The processing of fruit juice involves washing, extraction, clarification, and preservation.

Washing. Fruit is usually washed prior to any processing. Washing is typically conducted with a high-pressure soak or spray system. Under some conditions a surfactant or detergent may be added in order to release stubborn soil attached to the fruit. In apple processing a high-quality wash is necessary to ensure the safe removal of microorganisms responsible for mycotoxin formation and possible gastrointestinal poisoning.

Juice extraction. *Preparation.* Fruit is prepared for juice extraction by removing unwanted parts. This may include pitting operations for stone fruit such as apricots, cherries, or plums or peeling for such fruits as pineapples. In one large class of fruit, citrus fruit, juice extraction and separation from the peel are combined. Two major juice extraction systems for citrus exist. One is a reaming technique, in which the fruit is cut in half and the individual halves reamed to extract both the juice and the inner fruit solids. In the second major system, a hole is punched in the fruit and the juice squeezed out at the same time.

If the entire fruit is to be used in the juice, then typically it is disintegrated in a drum grater or a hammer mill. Care must be taken to control disintegration so that the particle size of the mash is compatible with the press system.

Rack-and-frame press

Pressing. Many different types of press are used for juice extraction. The most traditional is a rack-and-frame press, in which ground fruit (mash) is pumped into cloth partitions, called cheeses, which are separated by wooden or metallic racks. After a stack of cheeses has been produced, the press is activated and the juice expressed from the assembly.

Many variations of the rack-and-frame press exist. These include the continuous belt press, the bladder press, and the basket press.

Liquefaction. As an alternative to press systems, some processors have gone to total enzymatic liquefaction of the fruit mash. Cellulase and pectinase enzymes are added, and the mash is heated in order to accelerate the enzyme's performance.

Clarification. *Pectinization.* If the juice is to be clarified further or concentrated after extraction, treatment with pectinase may be required. The juice is monitored for pectin content using a qualitative pectin check, consisting of combining one part juice with two parts ethanol. If a gel forms, pectin is still present and depectinization must continue. When depectinization is complete, a floc is typically formed by the aggregation of partially degraded pectin-protein aggregates.

Filtration. Filtration systems are varied in design, operation, and application. The most traditional system is diatomaceous earth (DE) filtration, in which DE is used to aggregate and collect suspended solids. The DE is collected on filter paper inside the pressure filter as the juice passes through the unit. The resulting juice is sparkling clear. Owing to concern over the cost of DE and its disposal, other filtration processes have been designed. The most successful is membrane filtration, in which hollow fibre, open tubular, or ceramic membranes are employed in juice filtration systems.

Preservation. Once the juice has been clarified, it is ready to be preserved. In some cases large reserves of single-strength juice are kept in juice silos after having been pasteurized, but usually the juice is immediately processed into retail or institutional packages. For a single-strength juice packaging line, a typical process is to heat the juice to 88° C (190° F) and then bottle it. This produces a shelf-stable product.

For producing concentrate, the juice is passed through an evaporator, where the level of soluble solids is typically brought to 70 percent by weight. Retail packages of concentrate are typically filled at 45 percent dissolved solids; at this concentration a three-to-one dilution by the consumer will create a finished product with a soluble solid level of approximately 12 percent.

FRUIT PRESERVES, JAMS, AND JELLIES

The making of jellies and other preserves is an old and popular process, providing a means of keeping fruits far beyond their normal storage life and sometimes making use of blemished or off-grade fruits that may not be ideal for fresh consumption. In jelly making, the goal is to produce a clear, brilliant gel from the juice of a chosen fruit. Jams are made from the entire fruit, including the pulp, while preserves are essentially jellies that contain whole or large pieces. Marmalade, usually made from citrus fruit, is a jellylike concentrate of prepared juice and sliced peel.

Sugar, acid, and pectin

The essential ingredients for a successful preserve are sugar, acid, and pectin. These three ingredients lower the pH of the preserve and bind available water, thus creating an environment in which the growth of microorganisms is retarded. In some cases the fruit can provide all the pectin and acid that are needed. If the acid content of the fruit is low, external sources such as lemon juice can be added. Similarly, if the planned mix of fruit is low in pectin, a commercial source may be used. Sugar is always added, and in general all of the three essential ingredients have to be added in order to create a successful product.

The making of preserves begins with an initial mix containing not less than 45 parts by weight fruit for every 55 parts by weight sugar solids. The sugar solids are added after the fruit is crushed, and the mix is then cooked. Cooking may be done in a highly controlled vacuum kettle, in which flavour volatiles are captured and returned to the product. The cooking process continues until the heated mix is concentrated to a predetermined level of soluble solids. A generally accepted level is 65 percent soluble solids; at this concentration the boiling temperature is 7° to 12° above the boiling point of water. The product is then transferred to containers and sealed as a shelf-stable product.

The exact amount of sugar needed depends on the acidity level, the natural sugar content, and the type of product desired. If sugar content is too low, the resulting jelly will be tough; excessive sugar, on the other hand, will create a "soft set" that can be broken easily. Appropriate amounts of acid and pectin are added during the cooking process. The pH must be adjusted to an acidic level of approximately 3.1. Increased acidity reduces the amount of sugar needed in the blend, although excessive acidity can cause syneresis, or a separation of liquid from the gel.

If the pectin level is inadequate, then the preserve will not "set"; that is, not enough water will be bound to create a complete gel.

FRUIT PRESERVATION

Since fruits are generally acidic, they are naturally amenable to preservation. The premier role of acidity in preservation is to stop bacterial growth. Second, increased acidity can activate chemical reactions such as pectin set, which lowers water activity and reduces the possibility of microbial growth.

Dehydration. Dehydration is among the oldest and most common forms of fruit preservation. In dehydration, moisture in the fruit is driven off, leaving a stable food that has a moisture content below that at which microorganisms can grow. There are three basic systems for dehydration: sun drying, such as that used for raisins; hot-air dehydration; and freeze-drying.

Advantages

Dehydration has a number of advantages. Dehydrated fruit has a virtually unlimited shelf life when held under proper storage conditions. Drying does not significantly reduce the calories or minerals, and vitamin losses are similar to other preservation methods. In addition, by reducing the weight and the need for refrigeration, handling and transportation costs can be reduced dramatically. Dehydrated fruits are typically reduced in weight by 75 to 90 percent.

Encyclopædia Britannica, Inc.

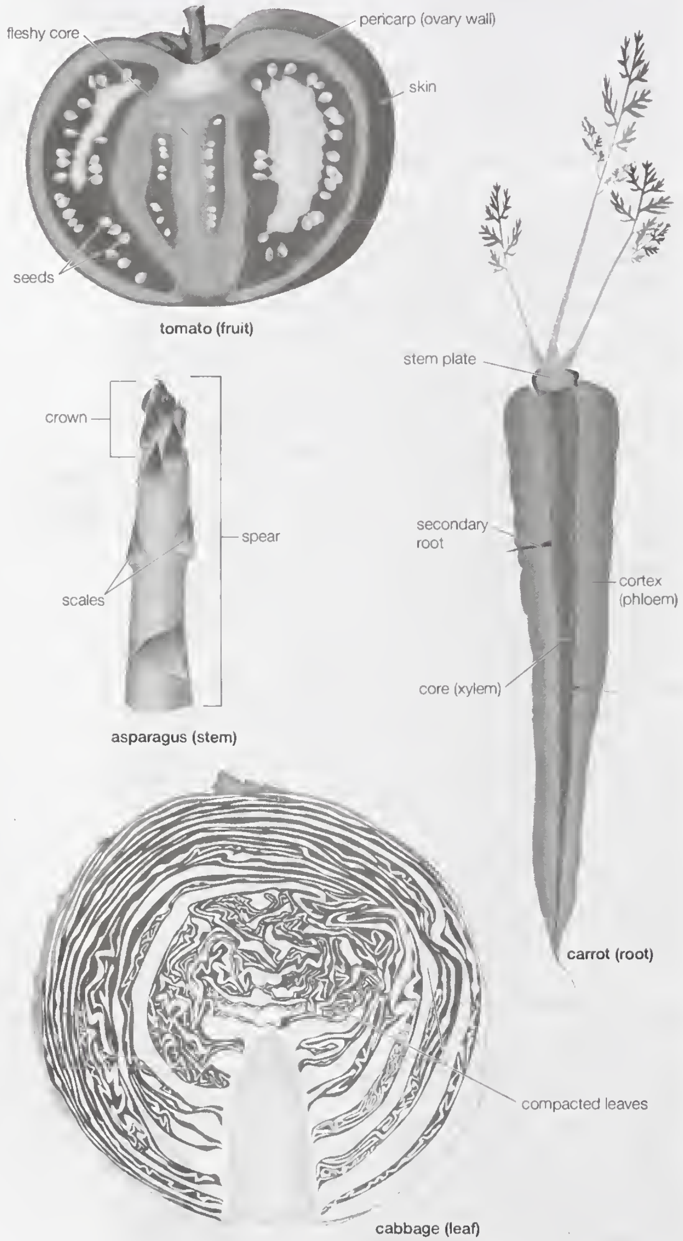


Figure 21: Structures of four representative vegetables.

Although dehydration offers a convenient product form, it usually requires a careful inactivation of enzymes. This is usually accomplished by blanching of the fruit or by chemical inactivation. Typically, sulfur dioxide is added for its antioxidant and preservative effects. In order to control browning, the fruit is often treated prior to dehydration with sodium sulfite and sodium bisulfite.

Thermal processes. In thermal processing, heat is used to destroy spoilage organisms and to inactivate troublesome enzymes. Enzymes are typically responsible for browning, softening, and the development of off-flavours. For high-acid fruit products the most typical thermal process is canning, in which fruit or fruit products are hot-filled or heated in a hermetically sealed container. The process temperature is generally in the range of 88° C (190° F).

Chemical preservation. Chemicals also can be used as a preservative, either through artificial addition or through the action of microorganisms. An example of the latter method is yeast fermentation, which can cause an increase in ethyl alcohol sufficient to preserve the fruit product. Pickling is another example of chemical preservation. In the case of pickling, the product may be preserved by the addition of salt, sugar, acetic acid (vinegar), and alcohol. High sugar content also acts as a fruit preservative by tying up all available moisture so that microorganisms cannot grow.

Irradiation. Although irradiation is an expensive method, it has been shown to be an effective means of extending the shelf life of fresh fruits. Irradiated fruit products have not been well received by the public, even in light of evidence supporting the healthfulness and safety of such foods.

Freezing. Freezing of fruits and fruit products is a common consumer practice. Cold temperatures act to retard the spoilage of fruit by inhibiting microbial action and slowing metabolic processes. In order to achieve extended storage life, the product must be held well below the freezing point of water—typically at a minimum of −23° C (−10° F). Generally, rapid freezing leads to an improved texture upon thawing.

A prerequisite for effective freezing is inactivation of fruit enzymes. Traditionally, this is done through blanching or by the addition of a chemical. Blanching consists of heating the fruit for a short time in water or steam prior to cooling and subsequent freezing. The blanch step is intended to inactivate enzyme systems responsible for off-flavours, browning, and softening. (M.R.McL.)

Vegetables

Vegetables consist of a large group of plants consumed as food. Perishable when fresh but able to be preserved by a number of processing methods, they are excellent sources of certain minerals and vitamins and are often the main source of dietary fibre. The consumption of vegetables has increased significantly as consumers have become more health-conscious. Owing to the perishable nature of the fresh produce, international trade in vegetables is mostly confined to the processed forms.

STRUCTURE AND COMPOSITION

Vegetables can be classified by edible parts into root (*e.g.,* potatoes and carrots), stem (asparagus and celery), leaf (lettuce and spinach), immature flower bud (broccoli and brussels sprouts), and fruit (tomatoes and cucumbers). The four basic types are illustrated in Figure 21.

Aging and spoilage. Depending on the class of vegetable, there are differences in the structure, size, shape, and rigidity of the individual cells. The fresh market shelf life and processing requirements are also very different. Vegetable cells, as plant cells, have rigid cell walls and are glued together by various polysaccharides such as cellulose, hemicellulose, and pectin. Once vegetables are harvested from the fields, the cells, now deprived of nutrient supplies normally obtained from soils and the air, go into senes-

Enzymatic reactions

cence, or aging. The most noticeable structural change in senescent vegetables is softening, or loss of texture. Softening is caused by natural enzymatic reactions that degrade the plant cell walls. A large group of enzymes is involved in the senescence stage, including cellulase, pectinase, hemicellulase, proteinase, and others. After these enzymes break open the cells, chemical oxidation reactions take place and the vegetables develop off-flavours and loss of nutritional value. Broken cells are also much more easily subject to microbial attacks, which quickly lead to spoilage. In addition, even though the vegetables may be packaged or bagged, the plant cells continue to respire, or break down carbohydrates for energy needs. Respiration leads to loss of quality, so that eventually the products are unsuitable for human consumption.

Nutritional value. The four quality factors of vegetables are colour, texture, flavour, and nutritive values. Fresh vegetables are purchased on the basis of colour and texture, but repeated purchases are made on the basis of flavour and nutritional content. The major nutrients contributed by vegetables to the human diet are dietary fibre (both soluble and insoluble), minerals (calcium, phosphorus, iron, sodium, potassium), and vitamins (vitamin C, vitamin A, thiamine, niacin, folic acid). The nutrient composition of selected vegetables is shown in Table 5.

Certain vegetables contribute lipids to the diet, mostly in the form of unsaturated oils. Roots and legumes can be important contributors of dietary proteins—especially in developing countries, where animal proteins are scarce. One potential nutritional problem of obtaining proteins from a single vegetable source is the low concentration of essential amino acids in vegetables. Twenty common amino acids are considered to be building blocks of proteins for the body. Of these 20, the body cannot synthesize eight; these eight must be obtained from foods. Most vegetable proteins are low in one of the eight essential amino acids; for example, corn is low in lysine, and soybeans are low in methionine. However, if proteins are obtained from a proper mixture of vegetables, there will not be a nutritional problem.

It is a common misconception that fresh vegetables are always superior in nutritional value to processed vegetables. Several investigations have shown frozen or canned vegetables can actually have higher nutritional value than fresh products. Fresh vegetables are subject to quality and vitamin losses during transportation and storage, whereas processing before these losses occur can yield a nutritionally superior product. Research has shown that a major cause of nutrient loss in vegetables is in the draining of cooking or processing liquids.

FRESH AND MINIMALLY PROCESSED VEGETABLES

Harvesting and storing. Most leafy vegetables that do not require harvesting by mechanical device are cooled immediately after harvest to remove field heat, sorted to remove debris, washed to remove dirt, and bundled or packed for shipping and retail. In most cases vegetables are bundled as whole plants, since cutting will injure the cells and liberate ethylene, which promotes senescence and shortens shelf life. Low-temperature storage is essential in the handling of quality leafy vegetables. On the other hand, storing below refrigerated temperature may lead to chilling injury of certain vegetables and to rapid loss of quality. In developing countries where refrigeration is not available, postharvest losses of fresh vegetables can be as much as half the total harvest.

For roots and legumes, the harvesting of which is normally done by machines, some sorting and grading are performed either in the field or at collection stations. Bulk handling of these vegetables is common, and few additional steps of preparation are performed before distribution. For vegetables that need to be stored for long periods of time, treatments to avoid microbial spoilage, insects, and small-animal invasion may be necessary. For some vegetables such as cucumbers, a washing and waxing step may be taken to improve the shelf life and the attractiveness of the produce.

Packaging. Provided in response to demands for convenient foods, minimally processed fresh produce has gained popularity in the marketplace. These vegetables go through additional preparation steps of washing, sorting, grading, cutting, and packaging into retail-size containers. In order to extend the shelf life of these products, vacuum packing and modified-atmosphere (MA) packaging are practiced. In most cases air is replaced by an atmosphere high in carbon dioxide and low in oxygen. This modified atmosphere can slow the respiration rate and therefore the senescence of cut vegetables. The most common products in American and European markets are various types of cut lettuces with shredded carrots, cabbages, and other vegetables. Modern packaging techniques employing "clean room" concepts make it possible for such vegetable products as salad mix and stir-fry mix to have shelf lives approaching those of the whole plants. The products can be shipped by refrigerated containers to overseas locations and still have a shelf life long enough to reach consumers.

Modified-atmosphere packaging

Minimally processed vegetables normally do not contain any preservatives and have not gone through any heat or chemical treatment. The disadvantage of these products is that refrigeration storage is essential, limiting its practice to developed countries.

PROCESSING OF VEGETABLES

Because of the varied growing and harvesting seasons of different vegetables at different locations, the availability of fresh vegetables differs greatly in different parts of the world. Processing can transform vegetables from perishable produce into stable foods with long shelf lives and thereby aid in the global transportation and distribution of many varieties of vegetables. The goal of processing is to deter microbial spoilage and natural physiological deterioration of the plant cells. Generally, the techniques include blanching, dehydrating, canning, freezing, fermenting and pickling, and irradiating.

Table 5: Nutrient Composition of Selected Vegetables and Vegetable Products (per 100 grams)*

vegetable or vegetable product	energy (kcal)	water (g)	carbohydrate (g)	vitamin C (mg)	thiamine (mg)	riboflavin (mg)	niacin (mg)	vitamin A (IU)	fat (g)	protein (g)
Asparagus, raw	23	92.40	4.54	13.2	0.140	0.128	1.170	583	0.20	2.28
Asparagus, canned	14	94.63	2.25	16.4	0.054	0.089	0.851	474	0.19	1.80
Cabbage, raw	25	92.15	5.43	32.2	0.050	0.040	0.300	133	0.27	1.44
Carrots, raw	43	87.79	10.14	9.3	0.097	0.059	0.928	28,129	0.19	1.03
Chinese cabbage, raw	13	95.32	2.18	45.0	0.040	0.070	0.500	3,000	0.20	1.50
Corn, sweet, raw	86	75.96	19.02	6.8	0.200	0.060	1.700	281	1.18	3.22
Corn, on the cob, frozen	98	71.79	23.50	7.2	0.103	0.088	1.681	246	0.78	3.28
Lettuce, iceberg, raw	13	95.89	2.09	3.9	0.046	0.030	0.187	330	0.19	1.01
Peas, green, raw	81	78.86	14.46	40.0	0.266	0.132	2.090	640	0.40	5.42
Peas, green, frozen	77	79.93	13.70	18.0	0.258	0.100	1.707	727	0.37	5.21
Potatoes, raw	79	78.96	17.98	19.7	0.088	0.035	1.484		0.10	2.07
Potatoes, mashed, dry flakes	354	6.51	81.21	83.6	1.031	0.110	6.146		0.39	8.35
Potato chips	536	1.90	52.90	31.1	1.167	0.197	3.827	0	34.60	7.00
Tomato juice, canned	17	93.90	4.23	18.3	0.047	0.031	0.673	556	0.06	0.76
Tomatoes, red, ripe	21	93.76	4.64	19.1	0.059	0.048	0.628	628	0.33	0.85
Tomatoes, sun-dried	258	14.56	55.76	39.2	0.528	0.489	9.050	874	2.97	14.11

*Values shown are approximations; actual nutrient composition can vary greatly depending on such factors as growing conditions, time of harvest, and storage.
Source: *Composition of Foods*, Agriculture Handbook no. 8-11, U.S. Department of Agriculture.

Blanching. After vegetables have been washed clean, they must undergo blanching (heating) in hot water at 88° C (190° F) for two to five minutes or with steam in a conveyor at 100° C (212° F) for one-half to one minute. Blanching inactivates natural enzymes that would cause discoloration and off-flavours and aromas. It also serves to reduce the number of microorganisms and to render vegetables limp for easy packing into containers. For some vegetables, such as spinach, snap beans, and collards, the blanching step also serves to remove harsh flavours.

After blanching the vegetables must go through rapid cooling in either cold water or cold air for better quality retention. The vegetables are then ready for the various food-processing methods described below.

Dehydration. Drying is probably the oldest method of preserving foods. The removal of water from vegetables is accomplished primarily by applying heat, whether it be through the radiant energy of the sun or through air heated by electrical energy. A major advantage of removing water is a reduction in volume and weight, which aids in storage and transportation of the dried products. Modern drying techniques are very sophisticated. Many machines are available to perform tunnel drying, vacuum drying, drum drying, spray drying, and freeze-drying. Although freeze-drying produces a food of outstanding quality, the cost is high, and it has not been used widely in vegetable products.

Dehydrated potatoes

One of the most familiar dehydrated products is instant potatoes. Almost all the mashed potato dishes served in restaurants and institutions are rehydrated instant potatoes. In restaurants and institutions dehydrated potato granules are used, while dehydrated flakes are preferred for home cooking. Potato granules have high bulk density and are easy to handle in large quantity. However, they produce mashed potatoes with a pasty texture—an effect caused by the rupture of cells during processing, so that starch is released from the cells. Mashed potatoes made from flakes, on the other hand, have a mealy texture comparable to that of freshly prepared mashed potatoes. The major difference in the processing of these two dehydrated products is in the drying steps. For granules, air-lift drying is used to bring the product to 10–13 percent moisture. After screening to proper granule size, the product is dried to 6 percent moisture in a fluidized-bed drier. In the making of flakes, a steam-heated drum drier is used to bring a flattened sheet of potato solids to final moisture content before it is broken into a suitable size for packaging. Although a considerable quantity of the potato cells are ruptured during the breaking of the dried sheet, the reconstituted product has an acceptably mealy texture because the potatoes are subjected to a precooking and cooling treatment as well as the addition of a monoglyceride emulsifier.

A small amount of sulfite may be used in producing certain dried vegetables. The sulfite serves as an antimicrobial agent, aids in heat transfer, and (in the case of potatoes) acts as a blanching agent. A small percentage of the consumer population is allergic to sulfite. Although the rehydrated product contains little or no sulfite, consumer concerns are forcing the industry to search for economically feasible sulfite replacements.

Storage of canned vegetables

Canning. Putting foods into metal cans or glass jars is the major food-processing method of the world. It is particularly useful in developing countries where refrigeration is limited or nonexistent. In the canning process, vegetables are often cut into pieces, packed in cans, and put through severe heat treatment to ensure the destruction of bacteria spores. The containers are sealed while hot so as to create a vacuum inside when they are cooled to room temperature. Properly processed canned vegetables can be stored at room temperature for ycars. Minor defects of the process, however, will result in bulged cans after long periods of storage. For safcty reasons, the contents of these cans should not be consumed. Although in most cases bulged cans are causcd by the formation of gas from chemical reactions between the mctal cans and their acidic contents, there is a remote possibility that inadequate heat processing did not destroy all bacteria spores. And, even though most heat-resistant spores are nonpathogenic, spores of *Clostridium botulinum* can survive undcrprocessing and produce deadly toxins that cause botulism.

Unfortunately, because of the severe heat treatment, some canned vegetables can have inferior quality and less nutritive value than fresh and frozen products. The nutrient most susceptible to destruction in canning is vitamin C.

For high-quality products, aseptic canning is practiced. Also known as high-temperature–short-time (HTST) processing, aseptic canning is a process whereby presterilized containers are filled with a sterilized and cooled product and sealed in a sterile atmosphere with a sterile cover. The process avoids the slow heat penetration inherent in the traditional in-container heating process, thus creating products of superior quality.

The canning process can be illustrated by the example of green beans (*Phaseolus vulgaris L.*). After arrival at the processing plant, the beans are conveyed to size graders. Graders consist of revolving cylinders with slots of various diameters through which the beans fall onto conveyers. The conveyers carry them to snipping machines, where their tips and stems are cut off. The snipped beans then pass over inspection belts, where defective beans are removed. Smaller beans are canned as whole beans, while larger beans are cut crosswise by machine into various lengths. Some smaller beans are cut lengthwise and marketed as French-cut beans. Both the small whole and cut beans are blanched for 1½ to 2 minutes in 82° C (150° F) water and mechanically packed in cans. The cans are then filled with hot water and dry salt or with brine, steam-exhausted for approximately five minutes, and sealed while hot or with steam flow. Depending on the size of the can, they are heat-processed for various periods of time—from 12 minutes at 120° C (250° F) to 36 minutes at 115° C (240° F). The cans are cooled to room temperature, labeled, and packaged for storage or immediate distribution.

Individually quick-frozen method

Freezing. Frozen foods have outstanding quality and nutritive value. Indeed, some frozen vegetables, such as green peas and sweet corn, may be superior in flavour to fresh produce. The high quality of frozen foods is mainly due to the development of a technology known as the individually quick-frozen (IQF) method. IQF is a method that does not allow large ice crystals to form in vegetable cells. Also, since each piece is individually frozen, particles do not cohere, and the final product is not frozen into a solid block. Various freezing techniques are commonly used in the preservation of vegetables. These include blast freezing, plate freezing, belt-tunnel freezing, fluidized-bed freezing, cryogenic freezing, and dehydrofreezing. The choice of method depends on the quality of end product desired, the kind of vegetable to be frozen, capital limitations, and whether or not the products are to be stored as bulk or as individual retail packages.

Most vegetables frozen commercially are intended for direct consumer use or for further processing into soups, prepared meals, or specialty items. Advances in packaging materials and techniques have led to bulk frozen products being stored in large retortable pouches. Many restaurants and institutions prefer bulk frozen soups packaged in these pouches because of their quality and convenience.

One of the most important vegetable crops preserved by freezing is sweet corn (*Zea mays L.*). Both corn on the cob and cut corn are frozen. Sweet corn must be harvested while still young and tender and while the kernels are full of "milk." After the ears are mechanically harvested, they are promptly hauled to the processing plant, where they are automatically dehusked and desilked. Probably more than any other vegetable, sweet corn loses its quality rapidly after harvest. Frozen corn maintains high quality by being processed within a few hours of picking. Corn on the cob is a particularly difficult vegetable to freeze. The dehusked and desilked ears are thoroughly washed and blanched in steam for 6 to 11 minutes and then promptly cooled. However, even an 11-minute blanch in steam does not completely inactivate all the enzymes in the cob portion. It is believed that the off-flavour frequently found in home-frozen corn on the cob comes from off-flavours produced in the cob that migrate out to the kernels. Blanched and cooled corn is quickly frozen by the fluidized-bed freezing process before packing. Blanched whole-kernel corn is produced either by blanching the corn on the cob before cutting; by partially blanching on the cob to set the milk, then cutting and blanching again; or by cutting before blanching. The "split" method of blanching twice produces the highest-quality product. After the corn is cut, impurities such as husk, silk, and imperfect kernels must be removed by either brine flotation or froth washing. In both methods the sound corn stays at the bottom while the impurities float off the tank. Whole-kernel corn can be frozen quickly using the individually quick-frozen method. Frozen corn can be packaged into polyethylene bags or cardboard cartons and labeled for retail, or it can be bulk-stored for further processing into components of value-added products such as frozen dinners.

Fermentation and pickling. In both fermented and pickled vegetables, acid is used to preserve the products. Pickled vegetables include cucumbers, green tomatoes, onions, radishes, and cabbages. The variety of vegetables used for fermentation or pickling may not be the same as fresh market vegetables. Owing to the acid environment, fermented or pickled vegetables need less heat treatment before being placed in containers.

Irradiation. Ionizing radiation, mostly gamma-ray, has been used in several countries to preserve vegetables. The practice is quite common in preventing potatoes from sprouting during long-term storage. Despite studies showing that products treated with low-dose ionizing radiation are safe, consumers are still concerned about this processing technology and have not accepted it. (Jo.J.J.)

Fish

Finfish, mollusks, and crustaceans

The word fish is commonly used to describe all forms of edible finfish, mollusks (*e.g.,* clams and oysters), and crustaceans (*e.g.,* crabs and lobsters) that inhabit an aquatic environment. Fish from the marine and freshwater bodies of the world have been a major source of food for humankind since before recorded history. Harvesting wild fish from fresh and marine waters and raising cultured fish in ponds were practices of ancient Egyptians, Greeks, and other Mediterranean peoples. Rudimentary processing techniques such as sun-drying, salting, and smoking were used by these ancient groups to stabilize the fish supply. Modern methods of processing and preservation have encouraged the consumption of many species of fish that are popular throughout the world.

CHARACTERISTICS OF FISH

Structure of skeletal muscles. The majority of edible fish products are derived from the skeletal muscles (flesh), which represent more than 50 percent of the total body mass of these animals. The skeletal muscles of fish differ from those of mammals and birds in that they are largely composed of stacks of short bundles of muscle fibres called myomeres. The myomeres are separated by thin horizontal (myosepta) and vertical (myocommata) layers of connective tissue. The unique structure and thin connective tissue sheaths of fish muscle give the meat its characteristic soft, flaky texture.

The skeletal muscles of fish are composed mostly of white, fast-twitch fibres. The high percentage of white fibres allows fish to swim with sudden, rapid movements and gives the meat its white colour. These fibres primarily metabolize glucose, a simple sugar released from muscle glycogen stores, for energy production through anaerobic (*i.e.,* in the absence of oxygen) glycolysis. Therefore, white fibres contain relatively little myoglobin, the oxygen-binding protein that provides the red colour of muscles in other animals.

Nutrient composition. The composition of fish may vary considerably—especially in their fat content—during certain growth periods and annual spawning or migration periods. In addition, the composition of fish bred in captivity (*i.e.,* aquaculture fish) may vary according to their artificial diet. Table 6 shows the nutrient composition of several types of fish.

Proteins. Fish are an excellent source of high-quality protein. Mollusks are generally lower in protein compared with finfish and crustaceans because of their high water content. The proteins found in fish are essentially the same as those found in the meat derived from other animals—that is, the sarcoplasmic proteins (*e.g.,* enzymes and myoglobin), the contractile or myofibrillar proteins (*e.g.,* actin and myosin), and the connective tissue proteins (*i.e.,* collagen).

Fat. The fat in fish is mostly liquid (*i.e.,* fish oil), because it contains a relatively low percentage of saturated fatty acids. Fish belong in a special nutritional class because they contain the omega-3 polyunsaturated fatty acids—eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA)—which have been shown to protect against several diseases, including heart disease. Unlike land plants, the marine and freshwater plants on which fish feed are rich in EPA and DHA.

Vitamins and minerals. Fish provide a number of important vitamins and minerals to the diet. They are a good source of the fat-soluble vitamins A, D, E, and K and the B vitamins riboflavin, niacin, and thiamine. The mineral content includes calcium, magnesium, phosphorus, and iron.

Microbial contamination

Microbiology. Because of their soft tissues and aquatic environment, fish are extremely susceptible to microbial contamination. At the time of harvest, fish carry a high microbial load on the surface of their skin, in their intestinal tract, and in their gills.

The type and number of microorganisms that live in fish vary according to the season, the species, and the natural habitat. Additional contamination may occur during the harvesting, handling, or processing of the fish. Common spoilage microorganisms of fish include species of *Pseudomonas, Moraxella,* and *Acinetobacter,* found mainly in marine fish, and *Bacillus* and *Micrococcus,* found in freshwater fish. Fish may also contain pathogenic (disease-causing) microorganisms such as *Salmonella* and *Escherichia coli.* Pathogenic contamination is of special

Table 6: Nutrient Composition of Raw Edible Portion of Fish Species (per 100 grams)

species	energy (kcal)	water (g)	protein (g)	fat (g)	cholesterol (mg)	calcium (mg)	iron (mg)	riboflavin (mg)	niacin (mg)
Catfish, channel (farmed)	135	75.38	15.55	7.59	47	9	0.50	0.075	2.304
Cod, Atlantic	82	81.22	17.81	0.67	43	16	0.38	0.065	2.063
Grouper, mixed species	92	79.22	19.38	1.02	37	27	0.89	0.005	0.313
Haddock	87	79.92	18.91	0.72	57	33	1.05	0.037	3.803
Halibut, Atlantic and Pacific	110	77.92	20.81	2.29	32	47	0.84	0.075	5.848
Herring, Atlantic	158	72.05	17.96	9.04	60	57	1.10	0.233	3.217
Mackerel, Atlantic	205	63.55	18.60	13.89	70	12	1.63	0.312	9.080
Salmon, Atlantic	142	68.50	19.84	6.34	55	12	0.80	0.380	7.860
Salmon, pink	116	76.35	19.94	3.45	52		0.77		
Trout, rainbow (wild)	119	71.87	20.48	3.46	59	67	0.70	0.105	5.384
Tuna, bluefin	144	68.09	23.33	4.90	38		1.02	0.251	8.654
Clam, mixed species	74	81.82	12.77	0.97	34	46	13.98	0.213	1.765
Crab, blue	87	79.02	18.06	1.08	78	89	0.74		
Lobster, northern	90	76.76	18.80	0.90	95			0.048	1.455
Oyster, Pacific	81	82.06	9.45	2.30		8	5.11	0.233	2.010
Scallop, mixed species	88	78.57	16.78	0.76	33	24	0.29	0.065	1.150
Shrimp, mixed species	106	75.86	20.31	1.73	152	52	2.41	0.034	2.552

Source: *Composition of Foods,* Agriculture Handbook no. 8-15, U.S. Department of Agriculture.

concern with mollusks because they are often eaten raw and as whole animals.

HANDLING OF HARVESTED FISH

The retention of nutritional properties and product quality of fish is dependent on proper handling of the catch after it has been harvested from its aquatic environment.

Chilling. Harvested fish must be immediately stored in a low-temperature environment such as ice or refrigerated seawater. This chilling process slows the growth of microorganisms that live in fish and inhibits the activity of enzymes. Because fish have a lower body temperature, softer texture, and less connective tissue than land animals, they are much more susceptible to microbial contamination and structural degradation. If immediate chilling is not possible, then the fish must generally be sold and eaten on the day of the harvest.

Ice cooling and holding normally requires a one-to-one or one-to-two weight ratio of ice to fish, depending on the specific geographic location and the time it takes to transport the fish to the processing plant. Refrigerated seawater cooling and holding causes less bruising and other structural damage to the fish carcasses than ice cooling. However, fish cooled in refrigerated seawater absorbs salt from the water. For this reason fish that is destined for sale on the fresh or frozen market may be held in refrigerated seawater for only a limited amount of time. The addition of salt during canning or smoking processes is adjusted in order to compensate for any absorbed salt.

Preprocessing. Preprocessing of fish prepares the raw material for final processing. It is often performed on shipboard or in a shore-based plant and includes such operations as inspection, washing, sorting, grading, and butchering of the harvested fish.

Butchering of fresh fish

The butchering of fish involves the removal of nonedible portions such as the viscera, head, tail, and fins (see Figure 22). Depending on the butchering process, as much as 30 to 70 percent of the fish may be discarded as waste or reduced to cheap animal feed. The lower figure applies when the fish is canned or sold as "whole." The higher figure applies when the fish is filleted or made into other pure meat products; in these cases the skeleton is discarded with as much as 50 percent of the edible flesh attached. Efforts to utilize this discarded fraction for the production of alternative food products have begun in the fish industry. (See below *Total utilization of raw materials.*)

FINAL PROCESSING OF FISH

The four basic procedures used in the final processing of fish products are heating, freezing, controlling water activity (by drying or adding chemicals), and irradiating. All these procedures increase the shelf life of the fish by inhibiting the mechanisms that promote spoilage and degradation. Each of these procedures also has an effect on the nutritional properties of the final product.

Heating. Heat treatment can significantly alter the quality and nutritional value of fish. Fish is exposed to heat during both the cooking process and the canning process.

Cooking. Fish is cooked in order to produce changes in the texture and flavour of the product and to kill pathogenic microorganisms. Heating fish to an internal temperature above 66° C or 150° F (*i.e.,* pasteurization conditions) is sufficient to kill the most resistant microorganisms. The cooking time must be closely regulated in order to prevent excessive loss of nutrients by heat degradation, oxidation, or leaching (the loss of water-soluble nutrients into the cooking liquid).

Canning. The canning process is a sterilization technique that kills microorganisms already present on the fish, prevents further microbial contamination, and inactivates degradative enzymes. In this process fish are hermetically sealed in containers and then heated to high temperatures for a given amount of time. Canned fish can be stored for several years. However, sterilization does not kill all microorganisms, and bacterial growth and gas production may occur if the products are stored at very high temperatures.

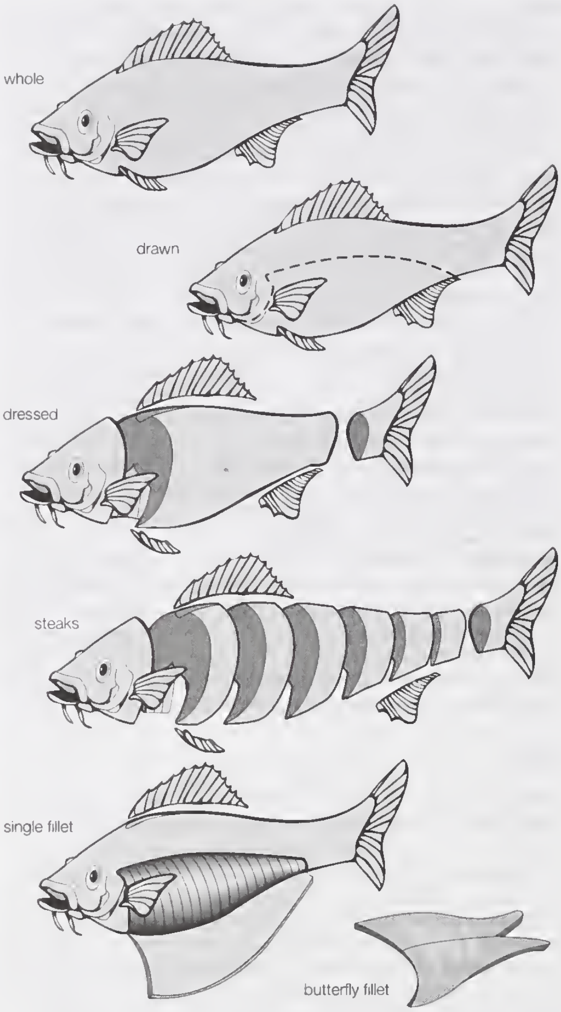


Figure 22: Schematic representation of various cuts of fish.
Encyclopædia Britannica, Inc

Because the severe thermal conditions of canning cause the disintegration and discoloration of the flesh of many species of fish, only a few types of fish are available as canned products. The most common types are tuna, salmon, herring, sardines, and shrimp. The thermal processing does not have a detrimental effect on the high-quality protein of the fish. In addition, these species are often canned with their bones left intact. The bones become soft and edible, significantly increasing the level of calcium present in the fish product. Tuna is an exception; because of special handling considerations, the bones of tuna are removed prior to canning. Tuna is normally caught far offshore and must be frozen and held for some period of time prior to canning. During this freezing and holding period unsaturated fatty acids are oxidized, causing the tuna to become rancid. The rancidity is removed by precooking, and the bones are removed at this time in order to facilitate the cutting and preparation of the meat for canning.

Canning of tuna

Freezing. Of the many processing methods used to preserve fish, only freezing can maintain the flavour and quality of fresh fish. Freezing greatly reduces or halts the biochemical reactions in fish flesh. For instance, in the absence of free water, enzymes cannot react to soften and degrade the flesh. The three steps for freezing fish include immediate cooling and holding, rapid freezing, and cold storage. If fish is frozen improperly, structural integrity may be compromised because of enzymatic degradation, texture changes, and dehydration.

Immediate cooling. The rapid cooling and holding of fish at temperatures between 2° and −2° C (36° and 28° F) takes place immediately after the fish have been harvested (see above *Handling of harvested fish: Chilling*).

Rapid freezing. The key to freezing is rapid reduction of the temperature to between −2° and −7° C (28° and 20° F). This temperature range represents the zone of maximum ice crystal formation in the cells of the flesh. If water in the cells freezes quickly, then the ice crystals will remain small and cause minimal damage to the cells. However, slow freezing results in the formation of large ice crystals and the rupturing of the cell membranes. When slow-frozen flesh is thawed, the ruptured cells release water (called drip) and many compounds that provide certain flavour characteristics of fish, resulting in a dry, tasteless product. Fish that passes through the zone of maximum ice crystal formation in less than one hour will generally have minimum drip loss upon thawing.

Cold storage. Once fish is frozen, it must be stored at a constant temperature of −23° C (−10° F) or below in order to maintain a long shelf life and ensure quality. A large portion of fresh fish is water (*e.g.,* oysters are more than 80 percent water). Because the water in fish contains many dissolved substances, it does not uniformly freeze at the freezing point of pure water. Instead, the free water in fish freezes over a wide range, beginning at approximately −2° C (28° F). The amount of remaining free water decreases until the product reaches a temperature of approximately −40° C (−40° F). Fish held below that temperature and packaged so as not to allow water loss through sublimation can be stored for an indefinite period. Unfortunately, there are relatively few commercial freezers capable of storing fish at −40° because of the tremendous variation in energy costs. Fish are therefore normally stored at −18° to −29° C (0° to −20° F), resulting in a variable shelf life ranging from a few weeks to almost one year.

Controlling water activity. Reducing the water activity of fish inhibits the growth of microorganisms and slows the chemical reactions that may be detrimental to the quality of the fish product. The control of water activity in fish is accomplished by drying, adding chemicals, or a combination of both methods.

Drying. The principal methods of drying, or dehydrating, fish are by forced-air drying, vacuum drying, or vacuum freeze-drying. Each of these methods involves adding heat to aid in the removal of water from the fish product. During the initial stages of drying, known as the constant-rate period, water is evaporated from the surface of the product and the temperature of the product remains constant. In the final stages of drying, known as the falling-rate period, the temperature of the product increases, causing water to move from the interior to the surface for evaporation.

Salt-curing

Curing. Curing reduces water activity through the addition of chemicals, such as salt, sugars, or acids. There are two main types of salt-curing used in the fish industry: dry salting and pickle-curing. In dry salting the butchered fish is split along the backbone and buried in salt (called a wet stack). Brine is drained off until the water content of the flesh is reduced to approximately 50 percent (the typical water content of fresh fish is 75 to 80 percent) and the salt content approaches 25 percent. In heavy or hard-cure salting, an additional step is taken in which warm air is forced over the surface of the fish until the water content is reduced to about 20 percent and the salt content is increased to approximately 30 percent. Most dry-salted fish products are consumed in warm, humid countries or in areas that have few means of holding products in refrigeration or cold storage.

In pickle-curing, fish are preserved in airtight barrels in a strong pickle solution formed by the dissolving of salt in the body fluids. This curing method is used for fatty fish such as herring.

Smoking. Traditionally, smoking was a combination of drying and adding chemicals from the smoke to the fish, thus preserving and adding flavour to the final product. However, much of the fish smoked today is exposed to smoke just long enough to provide the desired flavour with little, if any, drying. These products, called kippered fish, have short shelf lives, even under refrigeration, since the water activity remains high enough for spoilage organisms to grow.

The smoking process consists of soaking butchered fish in a 70 to 80 percent brine solution for a few hours to overnight, resulting in a 2 to 3 percent salt content in the fish. The fish are then partially dried on racks. As the brine on the surface dries, dissolved proteins produce a glossy appearance, which is one of the commercial criteria for quality. Smoking is carried out in kilns or forced-air smokehouses that expose the fish to smoke from smoldering wood or sawdust. In cold-smoking the temperature does not exceed 29° C (85° F), and the fish is not cooked during the process. Hot-smoking is more common and is designed to cook the fish as well as to smoke it.

Irradiating. Irradiation offers a means of pasteurizing or sterilizing a variety of food products. However, the use of this process has not been universally accepted throughout the food industry.

Food irradiators utilize radioisotopes, such as cobalt-60 (^{60}Co) or cesium-137 (^{137}Cs), or electron beam generators to provide a source of ionizing radiation. The irradiation of seafood has been extensively studied since the 1950s. The pasteurization of fresh fish using low-level dosages of ionizing radiation may extend the shelf life of the product up to several weeks. The sensory and nutritional characteristics of the fish are unaffected at these low levels of radiation.

TOTAL UTILIZATION OF RAW MATERIALS

In response to an increased demand for "ready-to-eat" fish products, along with a growing awareness of the limited supply of natural fish stocks, the fish industry has developed procedures for more efficient utilization of available raw materials. Because as much as 70 percent of harvested fish has traditionally been discarded or converted into cheap animal feeds, initial efforts to conserve fishery resources have focused on the development of edible products from underutilized species.

Gelling action of washed fish flesh

Surimi. *Surimi* was developed in Japan several centuries ago when it was discovered that washing minced fish flesh, followed by heating, resulted in a natural gelling of the flesh. When the *surimi* was combined with other ingredients, mixed or kneaded, and steamed, various fish gel products called *kamaboko* (fish cakes) were produced and sold as *neriseihin* (kneaded seafoods).

Modern *surimi* production consists of continuous operating lines with automated machinery for heading, gutting, and deboning of the fish; mincing, washing, and pressing (to remove water); and heating of the flesh. The *surimi* is then mixed with cryoprotectants and frozen for cold storage. Frozen *surimi* blocks are shipped to processing plants that produce various *kamaboko* products such as original *kamaboko* (*itatsuki*), broiled *kamaboko* (*chikuwa*), fried *kamaboko* (*satsumage*), and analog products, including imitation crab, scallops, and shrimp.

The chemistry of the *surimi* process involves the differential extraction of muscle proteins. The water-soluble sarcoplasmic proteins are removed during the washing of the minced flesh. These proteins inhibit the gelling properties of the minced flesh. The flesh is then comminuted with salt, which solubilizes the myofibrillar proteins actin and myosin. Upon heating, the myofibrillar proteins form a network structure that takes on a gellike consistency. Cryoprotectants are necessary to stabilize the myofibrillar protein network during frozen storage.

Minced fish flesh. The success of *surimi*-based products has stimulated the development of other products made from minced flesh. Minced fish products do not undergo the repeated washing cycles necessary for the production of *surimi.* Because of the presence of residual oils and sarcoplasmic enzymes (both oil and sarcoplasmic proteins are removed during the washing of *surimi*), cryoprotectants must also be added to the minced flesh prior to freezing in order to protect the product from oil oxidation and enzyme degradation.

Minced fish flesh is used in a wide variety of products. The largest volumes are extruded into formed patties for main dishes and sandwiches. The forming process involves combining the minced flesh with condiments and extruding the mix under pressure to produce the desired product, much like the formation of hamburger patties and sausages. The formed product may be battered and breaded in a final processing step. Other minced flesh products include nuggets and items used as hors d'oeuvres, fish chowders, and smoked fish sticks. (G.M.P.)

Meat

Meat is the common term used to describe the edible portion of animal tissues and any processed or manufactured products prepared from these tissues. Meats are often classified by the type of animal from which they are taken. Red meat refers to the meat taken from mammals; white meat refers to the meat taken from fowl; seafood refers to the meat taken from fish and shellfish; and game refers to meat taken from animals that are not commonly domesticated. In addition, most commonly consumed meats are specifically identified by the live animal from which they come. Beef refers to the meat from cattle, veal from calves, pork from hogs, lamb from young sheep, and mutton from sheep older than two years. It is with these latter types of red meat that this section is concerned.

CONVERSION OF MUSCLE TO MEAT

Muscle is the predominant component of most meat and meat products. Additional components include the connective tissue, fat (adipose tissue), nerves, and blood vessels that surround and are embedded within the muscles. The structural and biochemical properties of muscle are therefore critical factors that influence both the way animals are handled before, during, and after the slaughtering process and the quality of meat produced by the process.

Muscle structure and function. There are three distinct types of muscle in animals: smooth, cardiac, and skeletal. Smooth muscles, found in the organ systems including the digestive and reproductive tracts, are often used as casings for sausages. Cardiac muscles are located in the heart and are also often consumed as meat products. However, most meat and meat products are derived from skeletal muscles, which are usually attached to bones and, in the living animal, facilitate movement and support the weight of the body. Skeletal muscles are the focus of the following discussion.

Connective tissue

Skeletal muscle structure. Skeletal muscles are divided from one another by a covering of connective tissue called the epimysium. As shown in Figure 23, individual muscles are divided into separate sections (called muscle bundles) by another connective tissue sheath known as the perimysium. Clusters of fat cells, small blood vessels (capillaries), and nerve branches are found in the region between muscle bundles. Muscle bundles are further divided into smaller cylindrical muscle fibres (cells) of varying lengths that are individually wrapped with a thin connective tissue sheath called the endomysium. Each of the connective tissue sheaths found throughout skeletal muscle is composed of collagen, a structural protein that provides strength and support to the muscles.

The plasma membrane of a muscle cell, called the sarcolemma, separates the sarcoplasm (muscle cell cytoplasm) from the extracellular surroundings. Within the sarcoplasm of each individual muscle fibre are approximately 1,000 to 2,000 myofibrils. Composed of the contractile proteins actin and myosin, the myofibrils represent the smallest units of contraction in living muscle.

Skeletal muscle contraction. The contraction of skeletal muscles is an energy-requiring process. In order to perform the mechanical work of contraction, actin and myosin utilize the chemical energy of the molecule adenosine triphosphate (ATP). ATP is synthesized in muscle cells from the storage polysaccharide glycogen, a complex carbohydrate composed of hundreds of covalently linked molecules of glucose (a monosaccharide or simple carbohydrate). In a working muscle, glucose is released from the glycogen reserves and enters a metabolic pathway called glycolysis, a process in which glucose is broken down and the energy contained in its chemical bonds is harnessed for the synthesis of ATP. The net production of ATP depends on the level of oxygen reaching the muscle. In the absence of oxygen (anaerobic conditions), the products of glycolysis are converted to lactic acid, and relatively little ATP is produced. In the presence of oxygen (aerobic conditions), the products of glycolysis enter a second pathway, the citric acid cycle, and a large amount of ATP is synthesized by a process called oxidative phosphorylation.

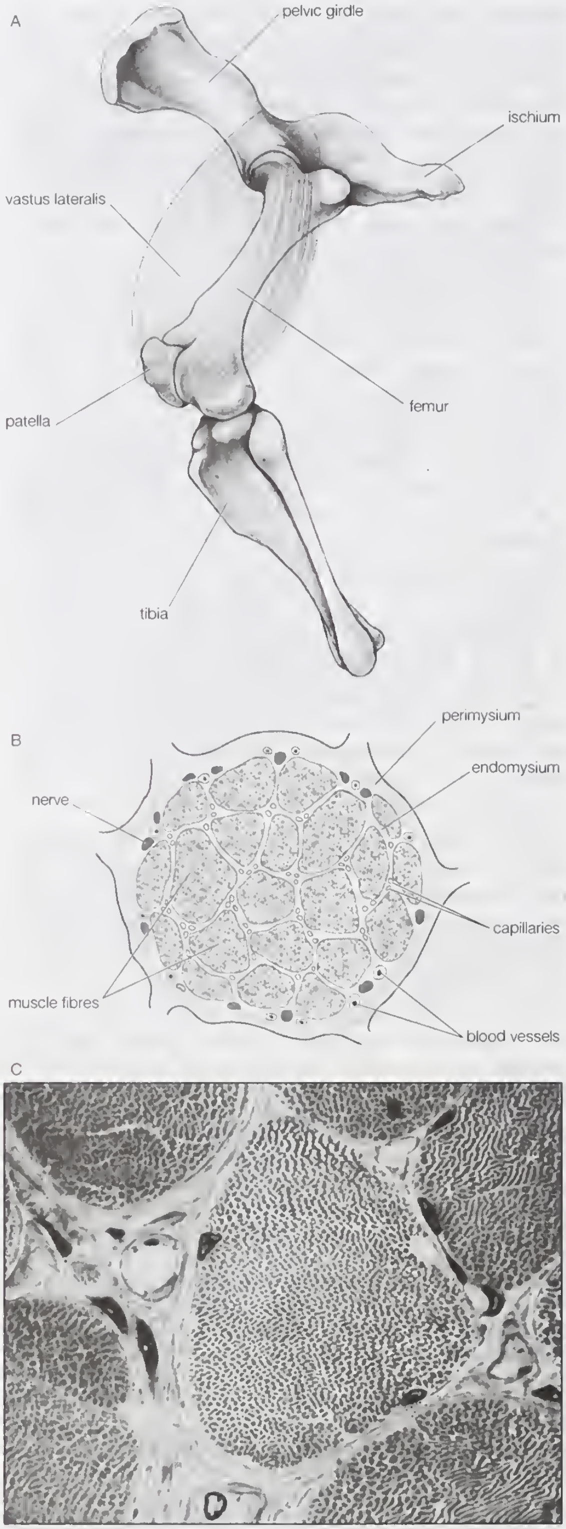


Figure 23: *Skeletal muscle structure.*
(A) Drawing of the vastus lateralis muscle of sheep.
(B) Cross section of a skeletal muscle. (C) Photomicrograph of muscle fibres.
(A,B) Encyclopædia Britannica, Inc., (C) © Ed Reschke/Peter Arnold, Inc

In addition to carbohydrates, fats supply a significant amount of energy for working muscles. Fats are stored in the body as triglycerides (also called triacylglycerols). A triglyceride is composed of three fatty acid molecules (nonpolar hydrocarbon chains with a polar carboxyl group at one end) bound to a single glycerol molecule. If the fat deposits are required for energy production, fatty acids are released from the triglyceride molecules in a process called fatty acid mobilization. The fatty acids are broken down into smaller molecules that can enter the citric acid cycle for the synthesis of ATP by oxidative phosphorylation.

Therefore, the utilization of fats for energy requires the presence of oxygen.

Myoglobin

An important protein of muscle cells is the oxygen-binding protein myoglobin. Myoglobin takes up oxygen from the blood (transported by the related oxygen-binding protein hemoglobin) and stores it in the muscle cells for oxidative metabolism. The structure of myoglobin includes a nonprotein group called the heme ring. The heme ring consists of a porphyrin molecule bound to an iron (Fe) atom. The iron atom is responsible for the binding of oxygen to myoglobin and has two possible oxidation states: the reduced, ferrous form (Fe^{2+}) and the oxidized, ferric form (Fe^{3+}). In the Fe^{2+} state iron is able to bind oxygen (and other molecules). However, oxidation of the iron atom to the Fe^{3+} state prevents oxygen binding.

Postmortem muscle. Once the life of an animal ends, the life-sustaining processes slowly cease, causing significant changes in the postmortem (after death) muscle. These changes represent the conversion of muscle to meat.

pH changes. Normally, after death, muscle becomes more acidic (pH decreases). When an animal is bled after slaughter (a process known as exsanguination), oxygen is no longer available to the muscle cells, and anaerobic glycolysis becomes the only means of energy production available. As a result, glycogen stores are completely converted to lactic acid, which then begins to build up, causing the pH to drop. Typically, the pH declines from a physiological pH of approximately 7.2 in living muscle to a postmortem pH of approximately 5.5 in meat (called the ultimate pH).

Protein changes. When the energy reserves are depleted, the myofibrillar proteins, actin and myosin, lose their extendability, and the muscles become stiff. This condition is commonly referred to as rigor mortis. The time an animal requires to enter rigor mortis is highly dependent on the species (for instance, cattle and sheep take longer than hogs), the chilling rate of the carcass from normal body temperature (the process is slower at lower temperatures), and the amount of stress the animal experiences before slaughter.

Eventually the stiffness in the muscle tissues begins to decrease owing to the enzymatic breakdown of structural proteins (*i.e.*, collagen) that hold muscle fibres together. This phenomenon is known as resolution of rigor and can continue for weeks after slaughter in a process referred to as aging of meat. This aging effect produces meats that are more tender and palatable.

PROPERTIES OF MEAT

Chemistry and nutrient composition. Regardless of the animal, lean muscle usually consists of approximately 21 percent protein, 73 percent water, 5 percent fat, and 1 percent ash (the mineral component of muscle). These figures vary as an animal is fed and fattened. Generally, as fat increases, the percentages of protein and water decrease. Table 7 provides a comparison of the nutrient composition of many meat products.

Protein. Meat is an excellent source of protein. As is explained above, these proteins carry out specific functions in living muscle tissue and in the conversion of muscle to meat. They include actin and myosin (myofibrillar proteins), glycolytic enzymes and myoglobin (sarcoplasmic proteins), and collagen (connective tissue proteins). Because the proteins found in meat provide all nine essential amino acids to the diet, meat is considered a complete source of protein.

Fat. Fats, in the form of triglycerides, accumulate in the fat cells found in and around the muscles of the animal. Fat deposits that surround the muscles are called adipose tissue, while fat that is deposited between the fibres of a muscle is called marbling.

In the diet the fats found in meat act as carriers for the fat-soluble vitamins (A, D, E, and K) and supply essential fatty acids (fatty acids not supplied by the body). In addition to their role as an energy reserve, fatty acids are precursors in the synthesis of phospholipids, the main structural molecules of all biological membranes.

Saturated and unsaturated fatty acids

Fatty acids are classified as being either saturated (lacking double bonds between their carbon atoms), monounsaturated (with one double bond), or polyunsaturated (containing several double bonds). The fatty acid composition of meats is dependent on several factors. In animals with simple stomachs, called nonruminants (*e.g.*, pigs), diet can significantly alter the fatty acid composition of meat. If nonruminants are fed diets high in unsaturated fats, the fat they deposit in their muscles will have elevated levels of unsaturated fatty acids. In animals with multichambered stomachs, called ruminants (*e.g.*, cattle and sheep), fatty acid composition found in the lean muscle is relatively unaffected by diet because microorganisms in the stomach alter the chemical composition of the fatty acids before they leave the digestive tract.

A beneficial characteristic of saturated fatty acids is that they do not undergo oxidation when exposed to air. However, the double bonds found in unsaturated fatty acids are susceptible to oxidation, and this oxidation promotes rancidity in meat. Therefore, products higher in saturated fats can generally be stored for a longer time without developing unpleasant flavours and odours.

Vitamins and minerals. Meat contains a number of essential vitamins and minerals. It is an excellent source of many of the B vitamins, including thiamine, choline, B_6, niacin, and folic acid. Some types of meat, especially liver, also contain vitamins A, D, E, and K.

Meat is an excellent source of the minerals iron, zinc, and phosphorus. It also contains a number of essential trace minerals, including copper, molybdenum, nickel, selenium, chromium, and fluorine. Table 7 provides a comparison of the vitamin and mineral content of different types of meat.

Table 7: Nutrient Composition of Cooked Lean Red Meats (per 100 grams)

meat type and cut	energy (kcal)	water (g)	protein (g)	fat (g)	cholesterol (mg)	vitamin B_{12} (μg)	thiamine (mg)	iron (mg)	zinc (mg)
Beef									
Chuck arm pot roast	219	58	33.02	8.70	101	3.40	0.080	3.79	8.66
Ribeye steak	225	59	28.04	11.70	80	3.32	0.100	2.57	6.99
Shortribs	295	50	30.76	18.13	93	3.46	0.065	3.36	7.80
Tenderloin	212	60	28.25	10.10	84	2.57	0.130	3.58	5.59
Top sirloin	200	61	30.37	7.80	89	2.85	0.130	3.36	6.52
Ground (extra lean)	265	54	28.58	15.80	99	2.56	0.070	2.77	6.43
Pork									
Loin roast	194	62	30.24	7.21	78	0.55	0.639	1.06	2.31
Tenderloin	164	66	28.14	4.81	79	0.55	0.940	1.47	2.63
Boston shoulder roast	232	61	24.21	14.30	85	0.93	0.669	1.56	4.23
Spareribs	397	40	29.06	30.30	121	1.08	0.382	1.85	4.60
Cured ham (extra lean)	145	68	20.93	5.53	53	0.65	0.754	1.48	2.88
Lamb									
Leg roast	191	64	28.30	7.74	89	2.64	0.110	2.12	4.94
Loin chop	202	63	26.59	9.76	87	2.16	0.100	2.44	4.06
Blade chop	209	63	24.61	11.57	87	2.74	0.090	2.07	6.48
Veal									
Loin chop	175	65	26.32	6.94	106	1.31	0.060	0.85	3.24
Rib chop	177	65	25.76	7.44	115	1.58	0.060	0.96	4.49

Source: *Composition of Foods*, Agriculture Handbook no. 8–10, 8–13, and 8–17, U.S. Department of Agriculture.

Cholesterol. Cholesterol is a constituent of cell membranes and is present in all animal tissues. Leaner meats typically are lower in cholesterol. Veal, however, is an exception: it is lower in fat than mature beef but has significantly higher cholesterol levels.

Carbohydrates. Meat contains virtually no carbohydrates. This is because the principal carbohydrate found in muscle, the complex sugar glycogen, is broken down in the conversion of muscle to meat (see above *Postmortem muscle: pH changes*). Liver is an exception, containing up to 8 percent carbohydrates.

Water. Water is the most abundant component of meat. However, because adipose tissue contains little or no moisture, as the percentage of fat increases in a meat cut, the percentage of water declines. Therefore, lean young veal may be as much as 80 percent water, while fully fattened beef may be as little as 50 percent. Because water is lost when meats are cooked, the percentages of protein and fat in cooked meats are usually higher than in their raw counterparts.

Colour. In well-bled animals approximately 80 to 90 percent of the total meat pigment is due to the oxygen-binding protein myoglobin. Colour differences in meat are related to the myoglobin content of muscle fibres and to the chemical state of the iron atom found in the myoglobin molecule.

White fibres and red fibres

Myoglobin content. A number of factors influence the myoglobin content of skeletal muscles. Muscles are a mixture of two different types of muscle fibre, fast-twitch and slow-twitch, which vary in proportions between muscles. Fast-twitch fibres have a low myoglobin content and are therefore also called white fibres. They are dependent on anaerobic glycolysis for energy production. Slow-twitch fibres have a high amount of myoglobin and a greater capacity for oxidative metabolism. These fibres are often called red fibres. Therefore, dark meat colour is a result of a relatively high concentration of slow-twitch fibres in the muscle of the animal.

A second factor contributing to the myoglobin content of a muscle is the age of the animal—muscles from older animals often have higher myoglobin concentrations. This accounts for the darker colour of beef relative to that of veal.

The size of an animal may also affect the myoglobin content of its muscles because of differences in basal metabolic rates (larger animals have a lower metabolism). Some smaller animals (such as rabbits) typically have a lower myoglobin concentration (0.02 percent of wet weight of muscle) and lighter coloured meat than larger animals such as horses (0.7 percent myoglobin) or deep-diving animals such as whales, which have very high concentrations of myoglobin (7 percent myoglobin) and dark, purple-coloured meat. Myoglobin concentration is also greater in intact males (animals that have not been castrated) of similar age, in muscles located closer to the bones, and in more physically active animals such as game.

Oxidation state of iron. The oxidation state of the iron atom of myoglobin also plays a significant role in meat colour. Meat such as beef viewed immediately after cutting is purple in colour because water is bound to the reduced iron atom of the myoglobin molecule (in this state the molecule is called deoxymyoglobin). Within 30 minutes after exposure to the air, beef slowly turns to a bright cherry-red colour in a process called blooming. Blooming is the result of oxygen binding to the iron atom (in this state the myoglobin molecule is called oxymyoglobin). After several days of exposure to air, the iron atom of myoglobin becomes oxidized and loses its ability to bind oxygen (the myoglobin molecule is now called metmyoglobin). In this oxidized condition, meat turns to a brown colour. Although the presence of this colour is not harmful, it is an indication that the meat is no longer fresh.

Tenderness. The tenderness of meat is influenced by a number of factors including the grain of the meat, the amount of connective tissue, and the amount of fat.

Meat grain. Meat grain is determined by the physical size of muscle bundles. Finer-grained meats are more tender and have smaller bundles, while coarser-grained meats are tougher and have larger bundles. Meat grain varies between muscles in the same animal and between the same muscle in different animals. As a muscle is used more frequently by an animal, the number of myofibrils in each muscle fibre increases, resulting in a thicker muscle bundle and a stronger (tougher) protein network. Therefore, the muscles from older animals and muscles of locomotion (muscles used for physical work) tend to produce coarser-grained meat.

Softening of collagen

Connective tissue. The amount of connective tissue in a muscle has a complex effect on the tenderness of the meat. The major component of connective tissue, collagen, has a tough, rigid structure. However, even though muscles from younger animals have more connective tissue, the meat derived from those muscles is generally more tender than that from older animals. This is due to the fact that collagen is broken down and denatured during the aging and cooking processes, forming a gelatin-like substance that makes the meat more tender. In addition, collagen becomes more rigid (resistant to breakdown and denaturation) with age, resulting in greater toughness of meat from older animals.

Fat. A high fat content within the adipose tissue and marbling sites of muscle contributes to the tenderness of the meat. During the cooking process the fat melts into a lubricant-type substance that spreads throughout the meat, increasing the tenderness of the final product.

LIVESTOCK SLAUGHTER PROCEDURES

The slaughter of livestock involves three distinct stages: preslaughter handling, stunning, and slaughtering. In the United States the humane treatment of animals during each of these stages is required by the Humane Slaughter Act. Figure 24 represents the general flow of the slaughter process.

Preslaughter handling. Preslaughter handling is a major concern to the livestock industry, especially the pork industry. Stress applied to livestock before slaughter can lead to undesirable effects on the meat produced from these animals. Preslaughter stress can be reduced by preventing the mixing of different groups of animals, by keeping livestock cool with adequate ventilation, and by avoiding overcrowding. Before slaughter, animals should be allowed access to water but held off feed for 12 to 24 hours to assure complete bleeding and ease of evisceration (the removal of internal organs).

Stunning. As the slaughter process begins, livestock are restrained in a chute that limits physical movement of the animal. Once restrained, the animal is stunned to ensure a humane end with no pain. Stunning also results in decreased stress of the animal and superior meat quality.

The three most common methods of stunning are mechanical, electrical, and carbon dioxide (CO_2) gas. The end result of each method is to render the animal unconscious. Mechanical stunning involves firing a bolt through the skull of the animal using a pneumatic device or pistol. Electrical stunning passes a current of electricity through the brain of the animal. CO_2 stunning exposes the animal to a mixture of CO_2 gas, which acts as an anesthetic.

Slaughtering. After stunning, animals are usually suspended by a hind limb and moved down a conveyor line for the slaughter procedures. They are typically bled (a process called sticking or exsanguination) by the insertion of a knife into the thoracic cavity and severance of the carotid artery and jugular vein. This method allows for maximal blood removal from the body. At this point in the process, the slaughtering procedures begin to differ by species.

Hogs. Hogs are usually stunned by electrical means or CO_2 gas. Mechanical stunning is not generally used in hogs because it may cause serious quality problems in the meat, including blood splashing (small, visible hemorrhages in the muscle tissue) in the lean meat.

Cleaning and dehairing

Hogs are one of the few domesticated livestock animals in which the skin is left on the carcass after the slaughter process. Therefore, after bleeding, the carcasses undergo an extensive cleaning procedure. First they are placed for about five minutes in a scalding tank of water that is between 57° and 63° C (135° and 145° F) in order to loosen hair and remove dirt and other material (called

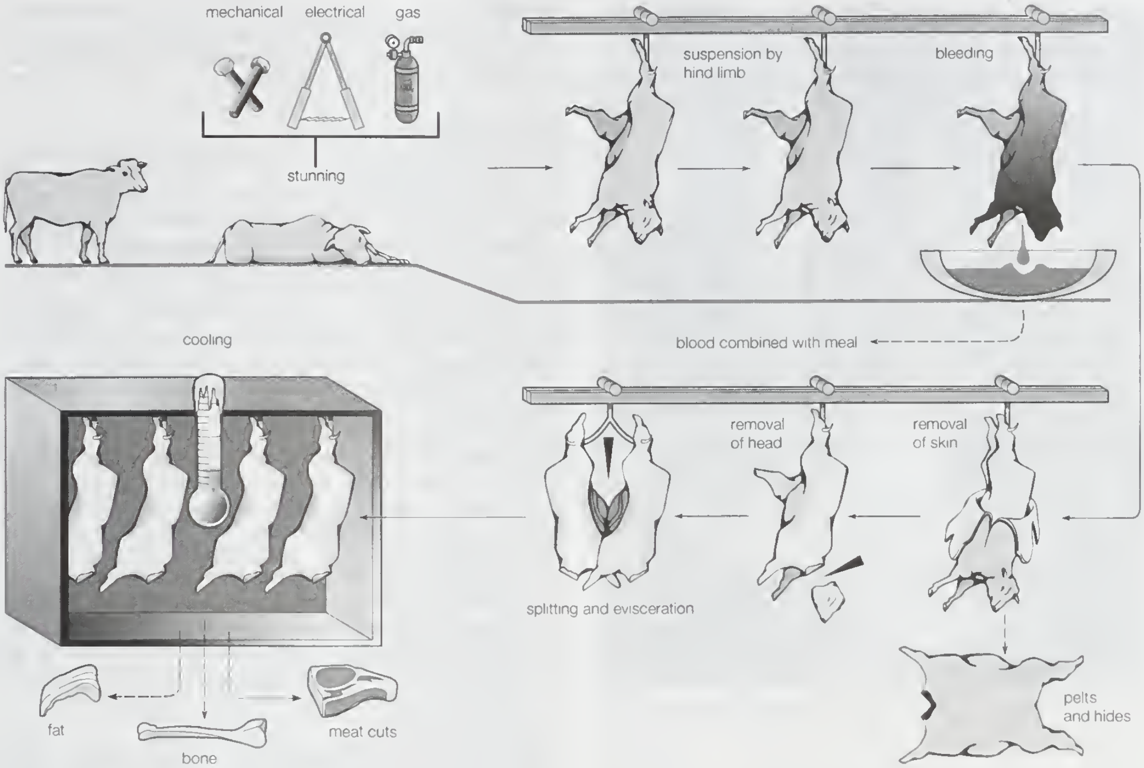


Figure 24: The basic slaughtering process.
Encyclopædia Britannica, Inc

scurf) from the skin. The careasses are then placed in a dehairing machine, whieh uses rubber paddles to remove the loosened hair. After dehairing, the careasses are suspended from a rail with hooks plaeed through the gambrel tendons on the hind limbs, and any residual hair is shaved and singed off the skin.

An exeeption to this proeedure oeeurs in certain specialized hog slaughter faeilities, sueh as "whole hog" sausage slaughter plants. In whole hog sausage production all the skeletal meat is trimmed off the earcass, and therefore the carcass is routinely skinned following exsanguination.

After eleaning and dehairing, heads are removed and earcasses are opened by a straight eut in the eentre of the belly to remove the viscera (the digestive system ineluding liver, stomaeh, bladder, and intestines and thc reproductive organs), pluek (thoraeie contents including heart and lungs), kidneys, and assoeiated fat (ealled leaf fat). The intestines are washed and eleaned to serve as natural casings for sausage produets. The earcasses are then split down the eentre of the baekbone into two "sides," whieh are placed in a cooler (called a "hot box") for approximately 24 hours before fabrieation into meat euts.

Cattle, calves, and sheep. These animals are usually stunned mechanically, but some sheep slaughter faeilities also use eleetrical stunning. The feet are removed from the careasses before they are suspended by the Achilles tendon of a hind leg for exsanguination. The carcasses are then skinned with the aid of mechanieal skinners ealled "hide pullers." Sheep pelts are often removed by hand in a process called "fisting." (In older operations, hides and pelts are removed by knife.) The hides (cattle and ealves) or pelts (sheep) are usually preserved by salting so that they can be tanned for leather produets. Heads are removed at the first eervical vertebra, ealled the atlas joint. Eviseeration and splitting are similar to hog proeedures, except that kidney, pelvie, and heart fat are typieally left in beef carcasses for grading. Careasses are then plaeed in a cooler for 24 hours (often 48 hours for beef) prior to fabrication into meat cuts.

By-products. By-products are the nonmeat materials colleeted during the slaughter proeess, commonly ealled offal. Variety meats include livers, brains, hearts, sweetbreads (thymus and pancreas), fries (testicles), kidneys, oxtails, tripe (stomaeh of eattle), and tongue. Bones and rendered meat are used as bone and meat mcal in animal feeds and fertilizers. Gelatin, obtained from high-eollagen produets sueh as pork snouts, pork skin, and dried rendered bone, is used in eonfeetions, jellies, and pharmaceuticals. Intestines are used as sausage easings. Hormones and other pharmaeeutieal produets sueh as insulin, heparin, and cortisone are obtained from various glands and tissues. Edible fats are used as lard (from hogs), tallow (from cattle), shortcnings, and cooking oils. Incdible fats are used in soap and candle manufaeturing and in various industrial grease formulations. Lanolin from sheep wool is used in cosmetics. Finally, hides and pelts are used in the manufaeture of leather.

MEAT INSPECTION

Meat inspeetion is mandatory and has the mission of assuring wholesomeness, safety, and aeeurate labeling of the meat supply. Although inspeetion procedures vary from eountry to country, they are eentred around the same basic prineiples and may be performed by government offieials, veterinarians, or plant personnel (see Figure 25). For example, in thc United States meat inspcction is administered through the Food Safety and Inspection Service of the United States Department of Agrieulture (USDA-FSIS) and is composed of several distinct programs. In general, these programs are representative of the basic inspeetion procedures used throughout the world and inelude antemortem inspeetion, postmortem inspeetion, reinspeetion during proeessing, sanitation, faeilities and equipment, labels and standards, eomplianee, pathology and epidemiology, residue monitoring and evaluation, federal-state relations, and foreign programs.

Inspeetion programs

Antemortem and postmortem inspeetion. Antemortem inspeetion identifies animals not fit for human eonsumption. Here animals that are down, disabled, diseased, or dead (known as 4D animals) are removed from the food ehain and labeled "eondemned." Other animals showing signs of being siek are labeled "suspeet" and are segregated from healthy animals for more thorough inspeetion during proeessing procedures.

Postmortem inspeetion of the head, viseera, and earcasses helps to identify whole carcasses, individual parts, or organs that are not wholesome or safe for human consumption.

Figure 25: Inspection of beef carcasses for grading by the U.S. Department of Agriculture.

© Jim Pickerell/Tony Stone Images

Reinspection during processing. Although previously inspected meat is used in the preparation of processed meat products, additional ingredients are added to processed meats. Reinspection during processing assures that only wholesome and safe ingredients are used in the manufacture of processed meat products (*e.g.*, sausage and ham).

Sanitation. Sanitation is maintained at all meat-packing and processing facilities by mandatory inspection both before and during the production process. This includes floors, walls, ceilings, personnel, clothing, coolers, drains, equipment, and other items that come in contact with food products. In addition, all water used in the production process must be potable (reasonably free of contamination).

Facilities and equipment. Facilities and equipment are inspected to ensure that they meet safety requirements. Facilities must have sufficient cooling and lighting, and rails from which carcasses are suspended must be high enough to assure that the carcasses never come in contact with the floor. Equipment must be able to be properly cleaned and must not adversely affect the wholesomeness of the products.

Labels and standards. Labels and standards regulations assure that products are accurately labeled, that nutritional information meets requirements, and that special label claims (*e.g.*, lean, light, natural) are accurate. Virtually all meat products must have the following components in their label: accurate product name, list of ingredients (in order of predominance), name and place of business of packer and manufacturer, net weight, inspection stamp and plant number, and handling instructions.

Compliance. Compliance assures that proper criminal, administrative, and civil sanctions are carried out against violators of food inspection laws. These violations include the sale of uninspected meat, the use of inaccurate labels, and the contamination of products.

Pathology and epidemiology. Pathology and epidemiology programs support the efforts of meat inspectors by working with other public health agencies to minimize the risk from widespread food-poisoning outbreaks. These agencies work to identify the causative agents of food poisoning and prevent repeated occurrences by improving prevention techniques (*e.g.*, proper handling and cooking and prevention of cross-contamination of raw and cooked products).

Residue monitoring and evaluation. Residue monitoring and evaluation programs identify animals containing harmful residues and remove them from the food chain. These residues include toxins from natural sources, from pesticides, from feeds, or from antibiotics administered to animals too soon before slaughter.

MEAT GRADING

Meat grading segregates meat into different classes based on expected eating quality (*e.g.*, appearance, tenderness, juiciness, and flavour) and expected yield of salable meat from a carcass. In contrast to meat-inspection procedures, meat-grading systems vary significantly throughout the world. These differences are due in large part to the fact that different countries have different meat quality standards. For example, in the United States cattle are raised primarily for the production of steaks and are fattened with high-quality grain feed in order to achieve a high amount of marbling throughout the muscles of the animal. High marbling levels are associated with meat cuts that are juicier, have more flavour, and are more tender. Therefore, greater marbling levels—and especially marbling that is finely textured and evenly distributed—improve the USDA quality grade (*i.e.*, Prime, Choice, or Select) of the beef. However, in Australia cattle are raised primarily for the production of ground beef products, and the highest quality grades are given to the leanest cuts of meat.

Some of the characteristics of meat used to assess quality and assign grades include: conformation of the carcass; thickness of external fat; colour, texture, and firmness of the lean meat; colour and shape of the bones; level of marbling; flank streaking; and degree of leanness.

RETAIL MEAT CUTTING

In the American style of meat cutting, whole carcasses are usually fabricated into more manageable primal (major) or subprimal (minor) cuts at the packing plant. This preliminary fabrication eases meat merchandising by reducing variability within the cuts. Primal and subprimal cuts are usually packaged and sold to retailers that further fabricate them into the products that are seen in the retail case.

Pork fabrication. Hogs are slaughtered at approximately 108 kilograms (240 pounds) and yield carcasses weighing approximately 76 kilograms (70 percent yield of live weight). Pork carcasses are usually divided into two sides before chilling, and each side is divided into four lean cuts plus other wholesale cuts. The four lean cuts are the ham, loin, Boston butt (Boston shoulder), and picnic shoulder. Figure 26 (top left) shows the major wholesale cuts of pork and the retail cuts derived from each.

Beef fabrication. Steers and heifers average 495 kilograms at slaughter and produce carcasses weighing 315 kilograms (63 percent yield of live weight). Beef carcasses are split into two sides on the slaughter floor. After chilling, each side is divided into quarters, the forequarter and hindquarter, between the 12th and 13th ribs. The major wholesale cuts fabricated from the forequarter are the chuck, brisket, foreshank, rib, and shortplate. The hindquarter produces the short loin, sirloin, rump, round, and flank. Figure 26 (top right) shows the major wholesale cuts of beef and the retail cuts derived from each.

Beef quarters

Lamb fabrication. Live sheep averaging 45 kilograms yield 22-kilogram carcasses (50 percent yield of live weight). Lamb carcasses are divided into two halves, the foresaddle and hindsaddle, on the fabrication floor. The foresaddle produces the major wholesale cuts of the neck, shoulder, rib, breast, and foreshank. The hindsaddle produces the major wholesale cuts of the loin, sirloin, leg, and hindshank. Figure 26 (bottom right) shows the major wholesale cuts of lamb and the retail cuts derived from each.

Veal fabrication. Veal is classified into several categories based on the ages of the animals at the time of slaughter. Baby veal (bob veal) is 2–3 days to 1 month of age and yields carcasses weighing 9 to 27 kilograms. Vealers are 4 to 12 weeks of age with carcasses weighing 36 to 68 kilograms. Calves are up to 20 weeks of age with carcasses ranging from 56 to 135 kilograms.

After slaughter, veal carcasses are split on the fabrication floor into two halves, the foresaddle and hindsaddle. The foresaddle produces the major wholesale cuts of the

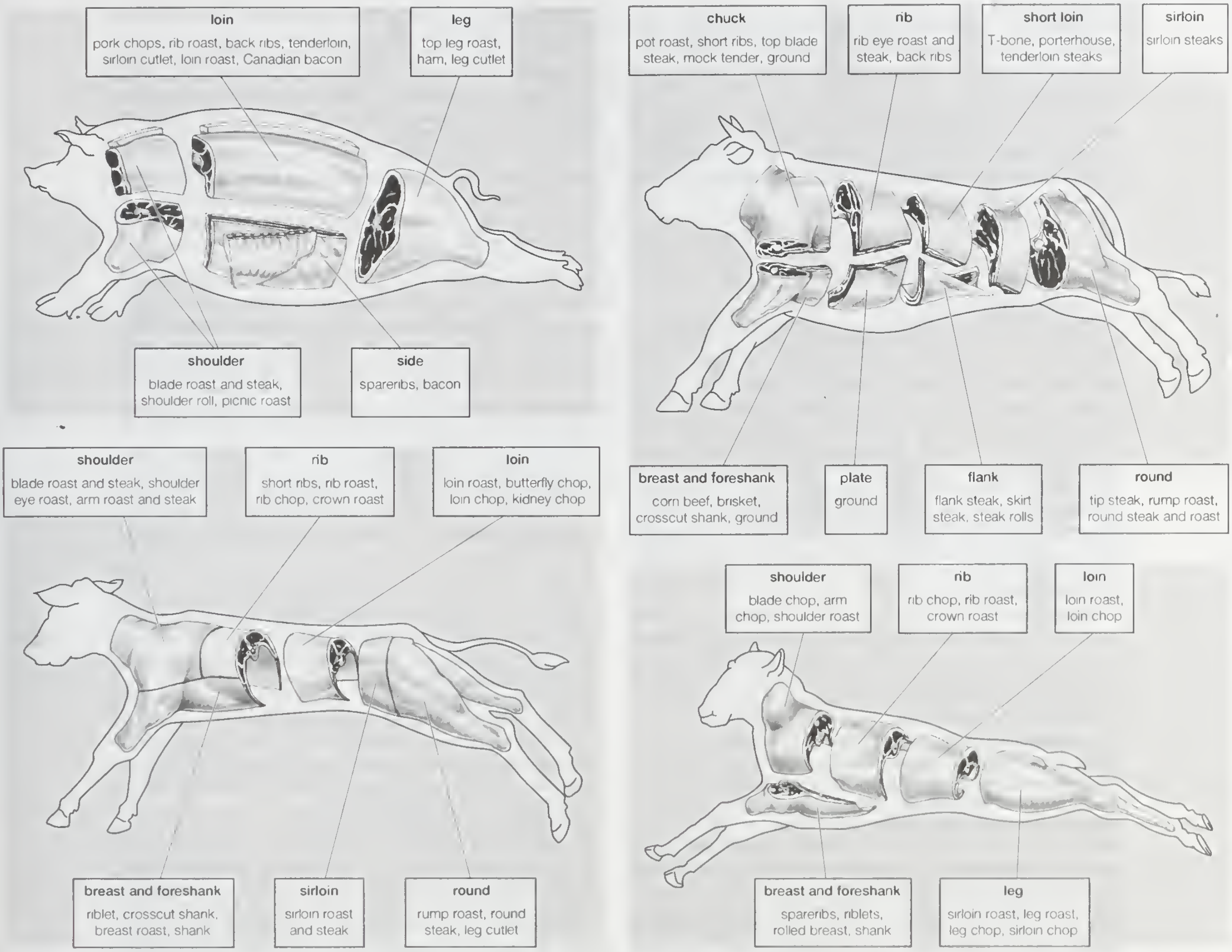

loin
pork chops, rib roast, back ribs, tenderloin, sirloin cutlet, loin roast, Canadian bacon
leg
top leg roast, ham, leg cutlet
shoulder
blade roast and steak, shoulder roll, picnic roast
side
spareribs, bacon
chuck
pot roast, short ribs, top blade steak, mock tender, ground
rib
rib eye roast and steak, back ribs
short loin
T-bone, porterhouse, tenderloin steaks
sirloin
sirloin steaks
breast and foreshank
corn beef, brisket, crosscut shank, ground
plate
ground
flank
flank steak, skirt steak, steak rolls
round
tip steak, rump roast, round steak and roast
shoulder
blade roast and steak, shoulder eye roast, arm roast and steak
rib
short ribs, rib roast, rib chop, crown roast
loin
loin roast, butterfly chop, loin chop, kidney chop
breast and foreshank
riblet, crosscut shank, breast roast, shank
sirloin
sirloin roast and steak
round
rump roast, round steak, leg cutlet
shoulder
blade chop, arm chop, shoulder roast
rib
rib chop, rib roast, crown roast
loin
loin roast, loin chop
breast and foreshank
spareribs, riblets, rolled breast, shank
leg
sirloin roast, leg roast, leg chop, sirloin chop

Figure 26: Wholesale and retail cuts of pork (top left), beef (top right), veal (lower left), and lamb (lower right).
Encyclopædia Britannica, Inc

shoulder, rib, breast, and shank. The hindsaddle produces the major wholesale cuts of the loin, sirloin, and round. Figure 26 (bottom left) shows the major wholesale cuts of veal and the retail cuts derived from each.

MEAT COOKERY

The physical changes associated with cooking meat are caused by the effects of heat on connective tissue and muscle proteins.

Colour changes. In beef, changes in cooking temperatures ranging from 54° C or 130° F (very rare) to 82° C or 180° F (very well done) correspond to changes in colour from deep red or purple to pale gray. These colour changes are a result of the denaturation of the myoglobin in meat. Denaturation is the physical unfolding of proteins in response to such influences as extreme heat. The denaturation of myoglobin makes the protein unable to bind oxygen, causing the colour to change from the bright cherry red of oxymyoglobin to the brown of denatured myoglobin (equivalent to metmyoglobin).

Effects of heat on fibres and connective tissues

Structural changes. The colour changes during cooking correspond to structural changes taking place in the meat. These structural changes are due to the effects of heat on collagen (connective tissue protein) and actin and myosin (myofibrillar proteins). In the temperature range between 50° and 71° C (122° to 160° F) connective tissue in the meat begins to shrink. Further heating to temperatures above 71° C causes the complete denaturation of collagen into a gelatin-like consistency. Therefore, tough meats with relatively high amounts of connective tissues can be slowly cooked under moist conditions to internal temperatures above 71° C and made tender by gelatinization of the collagen within the meat, while at the same time maintaining juiciness.

The myofibrillar proteins also experience major changes during cooking. In the range of 40° to 50° C (104° to 122° F) actin and myosin begin to lose solubility as heat denaturation begins. At temperatures of 66° to 77° C (150° to 170° F) the myofibrillar proteins begin to shorten and toughen. Beyond 77° C (170° F) proteins begin to lose structural integrity (*i.e.,* they are completely denatured) and tenderness begins to improve.

The effects of heat on both connective tissue and myofibrillar proteins must be balanced in order to achieve maximum tenderness during cooking. Meats with low amounts of connective tissue are most tender when served closer to medium rare or rare so that muscle proteins are not hardened. Conversely, meats with heavy amounts of connective tissue require slow cooking closer to well done in order to achieve collagen gelatinization.

MEAT MICROBIOLOGY, SAFETY, AND STORAGE

When the conversion of muscle to meat begins, biological degradation of meat also commences. In the absence of a living immune system, microorganisms are unchecked in their ability to grow and reproduce on meat surfaces.

Food-borne microorganisms. Generally, food-borne microorganisms can be classified as either food-spoilage or food-poisoning, with each presenting unique characteristics and challenges to meat product safety and quality.

Food-spoilage microorganisms. These organisms are responsible for detrimental quality changes in meat. The changes include discoloration, unpleasant odours, and physical alterations. The principal spoilage organisms are molds and bacteria.

Molds usually appear dry and fuzzy and are white or green in colour. They can impart a musty flavour to meat. Common molds in meat include the genera *Cladosporium, Mucor,* and *Alternaria*. Slime molds produce a soft, creamy material on the surface of meat.

Common spoilage bacteria include *Pseudomonas, Acinetobacter,* and *Moraxella.* Under anaerobic conditions, such as in canned meats, spoilage can include souring, putrefaction, and gas production. This is a result of anaerobic decomposition of proteins by the bacteria.

Food-poisoning microorganisms. Food-poisoning microorganisms can cause health problems by either intoxication or infection. Intoxication occurs when food-poisoning microorganisms produce a toxin that triggers sickness when ingested. Several different kinds of toxins are produced by the various microorganisms. These toxins usually affect the cells lining the intestinal wall, causing vomiting and diarrhea. Microorganisms capable of causing food-poisoning intoxication include *Clostridium perfringens* (found in temperature-abused cooked meats—*i.e.,* meats that have not been stored, cooked, or reheated at the appropriate temperatures), *Staphylococcus aureus* (found in cured meats), and *Clostridium botulinum* (found in canned meats).

Toxins

Infection occurs when an organism is ingested by the host, then grows inside the host and causes acute sickness and, in extreme cases, death. Common infectious bacteria capable of causing food poisoning in undercooked or contaminated meats are *Salmonella, Escherichia coli, Campylobacter jejuni,* and *Listeria monocytogenes.*

Preservation and storage. Meat preservation helps to control spoilage by inhibiting the growth of microorganisms, slowing enzymatic activity, and preventing the oxidation of fatty acids that promote rancidity. There are many factors affecting the length of time meat products can be stored while maintaining product safety and quality. The physical state of meat plays a role in the number of microorganisms that can grow on meat. For example, grinding meat increases the surface area, releases moisture and nutrients from the muscle fibres, and distributes surface microorganisms throughout the meat. Chemical properties of meat, such as pH and moisture content, affect the ability of microorganisms to grow on meat. Natural protective tissues (fat or skin) can prevent microbial contamination, dehydration, or other detrimental changes. Covering meats with paper or protective plastic films prevents excessive moisture loss and microbial contamination.

Cold storage. Temperature is the most important factor influencing bacterial growth. Pathogenic bacteria do not grow well in temperatures under 3° C (38° F). Therefore, meat should be stored at temperatures that are as cold as possible. Refrigerated storage is the most common method of meat preservation. The typical refrigerated storage life for fresh meats is 5 to 7 days.

Freezer storage is an excellent method of meat preservation. It is important to wrap frozen meats closely in packaging that limits air contact with the meat in order to prevent moisture loss during storage. The length of time meats are held at frozen storage also determines product quality. Under typical freezer storage of −18° C (0° F) beef can be stored for 6 to 12 months, lamb for 6 to 9 months, pork for 6 months, and sausage products for 2 months.

Freezing. The rate of freezing is very important in maintaining meat quality. Rapid freezing is superior; if meats are frozen slowly, large ice crystals form in the meat and rupture cell membranes. When this meat is thawed, much of the original moisture found in the meat is lost as purge (juices that flow from the meat). For this reason cryogenic freezing (the use of supercold substances such as liquid nitrogen) or other rapid methods of freezing meats are used at the commercial level to maintain maximal product quality. It is important to note, however, that freezing does not kill most microorganisms; they simply become dormant. When the meat is thawed, the spoilage continues where it left off.

Rapid freezing

Thawing meats often can cause more detrimental quality changes than freezing. In contrast to freezing, thawing should be a slow process. Meats are best thawed in the refrigerator with packaging left intact, so that moisture loss is minimized. Placing frozen meats out on a warm countertop or under warm water subjects the meat's outer layers to room temperatures for long periods of time before the meat is ready for cooking (completely thawed). This rapid method provides a conducive environment for the growth of food-borne microorganisms and increases the risk of food poisoning.

Vacuum packaging. Oxygen is required for many bacteria to grow. For this reason most meats are vacuum-packaged, which extends the storage life under refrigerated conditions to approximately 100 days. In addition, vacuum packaging minimizes the oxidation of unsaturated fatty acids and slows the development of rancid meat.

Canning. The second most common method of meat preservation is canning. Canning involves sealing meat in a container and then heating it to destroy all microorganisms capable of food spoilage. Under normal conditions canned products can safely be stored at room temperature indefinitely. However, certain quality concerns can compel processors or vendors to recommend an optimal "sell by" date.

Drying. Drying is another common method of meat preservation. Drying removes moisture from meat products so that microorganisms cannot grow. Dry sausages, freeze-dried meats, and jerky products are all examples of dried meats capable of being stored at room temperature without rapid spoilage.

Fermentation. One ancient form of food preservation used in the meat industry is fermentation. Fermentation involves the addition of certain harmless bacteria to meat. These fermenting bacteria produce acid as they grow, lowering the pH of the meat and inhibiting the growth of many pathogenic microorganisms.

Irradiation. Irradiation, or radurization, is a pasteurization method accomplished by exposing meat to doses of radiation. Radurization is as effective as heat pasteurization in killing food-spoilage microorganisms. Irradiation of meat is accomplished by exposing meat to high-energy ionizing radiation produced either by electron accelerators or by exposure to gamma-radiation-emitting substances such as cobalt-60 or cesium-137. Irradiated products are virtually identical in character to nonirradiated products, but they have significantly lower microbial contamination. Irradiated fresh meat products still require refrigeration and packaging to prevent spoilage, but the refrigerated storage life of these products is greatly extended.

Curing and smoking. Meat curing and smoking are two of the oldest methods of meat preservation. They not only improve the safety and shelf life of meat products but also enhance the colour and flavour. Smoking of meat decreases the available moisture on the surface of meat products, preventing microbial growth and spoilage. Meat curing, as commonly performed in products such as ham or sausage, involves the addition of mixtures containing salt, nitrite, and other preservatives.

Salt decreases the moisture in meats available to spoilage microorganisms. Nitrite prevents microorganisms from growing and retards rancidity in meats. Nitrite also produces the pink colour associated with cured products by binding (as nitric oxide) to myoglobin. However, the use of nitrite in meat products is controversial owing to its potential cancer-causing activity.

Sodium erythorbate or ascorbate is another common curing additive. It not only decreases the risks associated with the use of nitrite but also improves cured meat colour development. Other common additives include alkaline phosphates, which improve the juiciness of meat products by increasing their water-holding ability. (H.R.C.)

Poultry

Poultry is a major source of consumable animal protein. For example, per capita consumption of poultry in the

Table 8: Nutrient Composition of Roasted or Broiled Poultry Cuts (per 100 grams)

poultry type and cut	energy (kcal)	fat (g)	protein (g)	cholesterol (mg)	iron (mg)	zinc (mg)	vitamin B$_{12}$ (μg)	thiamine (mg)
Chicken								
Light meat with skin	222	10.85	29.02	84	1.14	1.23	0.32	0.060
Dark meat with skin	253	15.78	25.97	91	1.36	2.49	0.29	0.066
Light meat without skin	173	4.51	30.91	85	1.06	1.23	0.34	0.065
Dark meat without skin	205	9.73	27.37	93	1.33	2.80	0.32	0.073
Duck								
Flesh and skin	337	28.35	18.99	84	2.70	1.86	0.30	0.174
Flesh	201	11.20	23.48	89	2.70	2.60	0.40	0.260
Goose								
Flesh and skin	305	21.92	25.16	91	2.83	—	—	0.077
Flesh	238	12.67	28.97	96	2.87	—	—	0.092
Turkey								
Light meat with skin	197	8.33	28.57	76	1.41	2.04	0.35	0.056
Dark meat with skin	221	11.54	27.49	89	2.27	4.16	0.36	0.058
Light meat without skin	157	3.22	29.90	69	1.35	2.04	0.37	0.061
Dark meat without skin	187	7.22	28.57	85	2.33	4.46	0.37	0.063

Source: *Composition of Foods,* Agriculture Handbook no. 8–5, U.S. Department of Agriculture.

United States has more than quadrupled since the end of World War II as the industry has developed a highly efficient production system. Chickens and turkeys are the most common sources of poultry; however, other commercially available poultry meats come from ducks, geese, pigeons, quails, pheasants, ostriches, and emus.

CHARACTERISTICS OF POULTRY

Poultry is derived from the skeletal muscles of various birds and is a good source of protein, fat, and vitamins and minerals in the diet. (For a detailed discussion of muscles as food and their nutritional value, see above *Meat: Conversion of muscle to meat* and *Properties of meat.*) Table 8 shows the nutrient composition of several types of poultry.

Classification of birds. Birds bred for poultry production are generally grown for a particular amount of time or until they reach a specific weight. Rock Cornish hens, narrowly defined, are a hybrid cross specifically bred to produce small roasters; in the marketplace, however, the term is used to denote a small bird, five to six weeks old, that is often served whole and stuffed. Seven-week-old chickens are classified as broilers or fryers, and those that are 14 weeks old as roasters.

Fat content. The fat content of poultry differs in several ways from that found in red meat. Poultry has a higher proportion of unsaturated fatty acids compared with saturated fatty acids. Both turkey and chicken contain about 30 percent saturated, 43 percent monounsaturated, and 22 percent polyunsaturated fatty acids. The high levels of unsaturated fatty acids make poultry more susceptible to rancidity through the oxidation of the double bonds in the unsaturated fatty acids. Saturated fatty acids, on the other hand, do not contain double bonds in their hydrocarbon chains and are resistant to oxidation. However, this fatty acid ratio has led to the suggestion that poultry may be a more healthful alternative to red meat.

Location of fat deposits

In birds fat is primarily deposited under the skin or in the abdominal cavity. Therefore, a significant amount of the fat can be removed from poultry by removing the skin before eating.

Microbial contamination. Poultry provides an excellent medium for the growth of microorganisms. The principal spoilage bacteria found on poultry include *Pseudomonas, Staphylococcus, Micrococcus, Acinetobacter,* and *Moraxella.* In addition, poultry often supports the growth of certain pathogenic (disease-causing) bacteria, such as *Salmonella.*

Potential causes of contamination of poultry during the slaughtering and processing procedures include contact of the carcass with body parts that contain a high microbial load (*e.g.,* feathers, feet, intestinal contents), use of contaminated equipment, and physical manipulation of the meat (*e.g.,* deboning, grinding). Prevention of microbial contamination involves careful regulation and monitoring of the slaughtering and processing plants, proper handling and storage, and adequate cooking of raw and processed poultry products.

SLAUGHTERING PROCEDURES

Preslaughter handling. When the birds have reached "harvest" time, they are generally taken off of feed and water. This allows their digestive tracts to empty and reduces the potential for contamination during processing.

At night the birds are caught by specially trained crews and placed into plastic or wooden transport cages. The birds are then transported to the slaughterhouse, where the trucks are often kept between sets of fans to ventilate the cages.

In the next step the birds are removed from the cages and transferred to continuously moving shackles where they are suspended by both legs. The transfer is often done in a dark room illuminated by a red light; the birds are not sensitive to the red light and this helps to keep them calm.

The handling and transfer of birds both on the farm and at the slaughterhouse can be stressful. Stress can have negative effects on the quality of the final meat product, and therefore efforts are constantly being made to improve the preslaughter processes.

Slaughtering. *Stunning and killing.* After the birds have been transferred to the moving shackles, they are usually stunned by running their heads through a water bath that conducts an electric current. Stunning produces unconsciousness, but it does not kill the birds. The birds are killed either by hand or by a mechanical rotary knife that cuts the jugular veins and the carotid arteries at the neck. Any birds not killed by the machine are quickly killed by a person with a knife assigned to the bleed area. The birds are permitted to bleed for a fixed amount of time, depending on size and species (*e.g.,* 1½ minutes for broilers). Any bird that is not properly bled will be noticeably redder after feather removal and will be condemned.

Scalding. Following bleeding, the birds go through scalding tanks. These tanks contain hot water that softens the skin so that the feathers can be removed. The temperature of the water is carefully controlled. If retention of the yellow skin colour is desired, a soft-scald is used (about 50° C or 122° F). If a white bird is desired, a higher scald temperature is used, resulting in the removal of the yellow pellicle. Turkeys and spent hens (egg-laying birds that have finished their laying cycles) are generally run at higher temperatures—59° to 60° C (138° to 140° F).

Defeathering. The carcasses then go through the feather-picking machines, which are equipped with rubber "fingers" specifically designed to beat off the feathers. The carcasses are moved through a sequence of machines, each optimized for removing different sets of feathers. At this point the carcasses are usually singed by passing through a flame that burns off any remaining feathers.

Wax dipping

An extra process, called wax dipping, is often used for waterfowl, since their feathers are more difficult to remove. Following the mechanical feather picking, the carcasses are dipped in a melted, dark-coloured wax. The wax is allowed to harden and then is peeled away, pulling out the feathers at the same time. The wax is reheated and the feathers are filtered out so that the wax can be reused. This process is usually performed twice.

Figure 27: Packaging of poultry.

© Charles Gupton/Tony Stone Images

The blood and feathers accumulated during these early steps are generally collected and rendered to make blood meal and feather meal. The feathers from ducks and geese are often carefully collected and used for down production.

Removal of heads and legs. The heads of the birds go into a channel where they are pulled off mechanically; the legs of the birds are removed with a rotary knife (much like a meat slicer) either at the hock or slightly below it, depending on national custom. The carcasses drop off the shackle and are rehung by their hock onto the eviscerating shackle line. By law in the United States, the scalding and defeathering steps must be separated by a wall from the evisceration steps in order to minimize cross-contamination.

Evisceration and inspection. At this point the preen, or oil, gland is removed from the tail and the vent is opened so that the viscera (internal organs) can be removed. Evisceration can be done either by hand (with knives) or by using complex, fully automated mechanical devices. Automated evisceration lines can operate at a rate of about 70 birds per minute. The equipment is cleaned (with relatively high levels of chlorine) after each bird.

The carcasses are generally inspected during the evisceration process. The inspection procedures in the poultry industry vary around the world and may be performed by government inspectors, veterinarians, or plant personnel, depending on a country's laws. For example, in the United States the viscera are removed and placed on the side of the bird. Inspectors from the U.S. Department of Agriculture then examine the entire bird. The plant provides each inspector with an assistant who carries out any adjustments required by the inspector (*e.g.,* removing the entire bird or removing some part of the bird). The rejected parts are placed in a container marked "inedibles," and the contents are generally dyed (often a blue-purple), under supervision of the inspector, in order to prevent possible mixing with edible parts.

Following inspection, the carcasses are further cleaned. The viscera are separated from the carcasses, and the edible offal are removed from the inedible offal. The heart, stomach, and liver are all considered edible offal and are independently processed. Stomachs are generally cut open and the inside yellow lining of the stomach along with the stomach contents are removed.

The lungs and kidneys are removed separately from the other visceral organs using a vacuum pipe. A final inspection is often carried out at this point, and the carcasses are then washed thoroughly.

Chilling. After the carcasses have been washed, they are chilled to a temperature below 4° C (40° F). The two main methods for chilling poultry are water chilling and air chilling.

Water chilling. Water chilling is used throughout North America and involves a prechilling step in which a countercurrent flow of cold water is used to lower the temperature of the carcasses. The carcasses are then moved into a chiller—a large tank specifically designed to move the carcasses through in a specific amount of time. Two tanks are used to minimize cross-contamination.

Water chilling leads to an increase in poultry weight, and the amount of water gained is carefully regulated. In the United States the legal limits for water pickup are 8 percent for birds going directly to market and 12 percent for birds that will be further processed (the assumption is that they will lose the extra 4 percent by the time they reach the consumer).

Air chilling. Air chilling is the standard in Europe. The carcasses are hung by shackles and moved through coolers with rapidly moving air. The process is less energy-efficient than water chilling, and the birds lose weight because of dehydration. Air chilling prevents cross-contamination between birds. However, if a single bird contains a high number of pathogens, this pathogen count will remain on the bird. Thus, water chilling may actually result in a lower overall bacterial load, because many of the pathogens are discarded in the water.

The final temperature of the carcasses before shipment is usually about −2° to −1° C (28° to 30° F), just above the freezing point for poultry. In some cases a slight crusting on the surface occurs during the final chilling. For water-chilled carcasses this final chilling takes place after packaging, when the carcasses are placed in an air chiller.

PROCESSING OF POULTRY

Raw poultry products. Whole or individual parts of birds may be packaged raw for direct sale. Poultry packaged in the United States must include instructions about safe handling, including the need to wash all equipment that has come in contact with raw poultry and the need to wash one's hands before preparing other foods. Most raw turkey is sold frozen, while most chicken is sold fresh.

Fresh poultry. The birds are generally cut into a number of pieces, which are placed on plastic foam trays and covered with a plastic film (see Figure 27). A "diaper" (absorbent paper with a plastic backing) is often used to catch any liquid that may be released from the meat. Fresh poultry should be used within 14 to 21 days after slaughter and generally should not be kept in the home refrigerator for more than three days. In the United States, poultry that has been frozen to a temperature of −5° to −4° C (22° to 24° F) and then allowed to thaw can legally be sold as "fresh."

Storage of fresh poultry

Frozen poultry. Most frozen poultry is vacuum-packed in plastic bags and then frozen in high-velocity freezers. The birds are kept in cold storage until needed. Before freezing, poultry may be injected with various salts, flavourings, and oils in order to increase the juiciness of the meat. Injections are usually done with a multi-needle automatic injector, and information about the added ingredients is indicated on the package label.

Frozen storage time (including poultry bought fresh and frozen in a home freezer) depends on the temperature of the freezer, the quality of the packaging, and the cycling of the freezer. For best results poultry should be used within three months. Frozen poultry products can be used directly in the frozen state or thawed first. Thawing should be done in the refrigerator or under running cold water to minimize the potential for microbial contamination.

Processed poultry products. Poultry may be further processed into other products. The number of processed poultry products has increased dramatically since the 1970s because of the low cost of poultry and its versatile, bland flavour.

Battering and breading. Some poultry products are battered (*e.g.,* with beer batter) or battered and breaded (*e.g.,* with cracker meal, bread crumbs, or cornmeal) for frying. The meat may be either cooked or raw prior to coating. For battered and breaded poultry, the pieces are passed through a flour-based batter containing leavening and then through the breading ingredients. Many types of baked breadings have been developed to meet different tastes (*e.g.,* Cajun or Japanese). To hold the breading to the poultry, the product is deep-fried for a short time. If the poultry is fully cooked in this process, the consumer will only have to heat the product before eating it. Chicken nuggets are a battered and breaded product that is marinated before coating.

Tumbling and massaging. In the manufacturing of many poultry products, the meat is mixed with a variety of nonmeat ingredients, including flavourings, spices, and salt. Tumbling and massaging are gentle methods that produce a uniform meat mixture. A tumbler is a slowly rotating drum that works the meat into a smooth mixture. A massager is a large mixing chamber that contains a number of internal paddles. Cured turkey products (*i.e.,* treated with sodium nitrite), such as turkey ham and turkey pastrami, are often tumbled or massaged during processing.

Smoking. Poultry may be smoked. Prior to smoking, the birds must be brined (soaked in a salt solution containing certain flavourings) and then allowed to dry. Smoking can be done using real wood shavings or a smoke flavouring. In the last case this must be labeled in the United States as "natural smoke flavor added."

Minced poultry products

Deboning and grinding. Further processed poultry products leave the backs, necks, and bones available for their own processing. These materials are run through a machine called a mechanical deboner or a meat-bone separator. In general, the crushed meat and bones are continuously pressed against a screen and the edible, soft materials pushed through the screen. The resulting minced product is similar in texture to ground beef and has been used for many poultry products such as frankfurters (hot dogs) and bologna. Poultry frankfurters and bologna are made using a process similar to that for beef and pork. The meat is combined with water or ice, salt, and seasonings and chopped to emulsify the materials. The mixture is stuffed into plastic casings and cooked in a smokehouse. The meat is then quickly chilled, peeled, and vacuum-packaged. Bologna is stuffed into a larger casing and is not necessarily peeled. (J.M.Re.)

Eggs

While the primary role of the egg obviously is to reproduce the species, most eggs laid by domestic fowl, except those specifically set aside for hatching, are not fertilized but are sold mainly for human consumption. Eggs produced in quantity come from chickens, ducks, geese, turkeys, guinea fowl, pigeons, pheasants, and quail. This section describes the processing of chicken eggs, which represent the bulk of egg production in the United States and Europe. Duck eggs are consumed as food in parts of Europe and Asia, and goose eggs are also a food in many European countries. Commercial production of turkey and pigeon eggs is almost entirely confined to those used for producing turkey poults and young pigeons (squabs). Pheasant and quail eggs provide birds for hobby or sport use.

CHARACTERISTICS OF THE EGG

Structure and composition. The structural components of the egg are shown in Figure 28. They include the shell and shell membranes (10 percent); the albumen or white (60 percent), including the thick albumen, the outer thin albumen, the inner thin albumen, and the chalazae; and the yolk (30 percent). In a fertilized egg the yolk supplies the nutrients and the albumen supplies the water necessary for the development of the embryo. In addition, the layers of albumen act as a cushion to protect the embryo from jarring movements, while the chalazae help to maintain the orientation of the embryo within the egg.

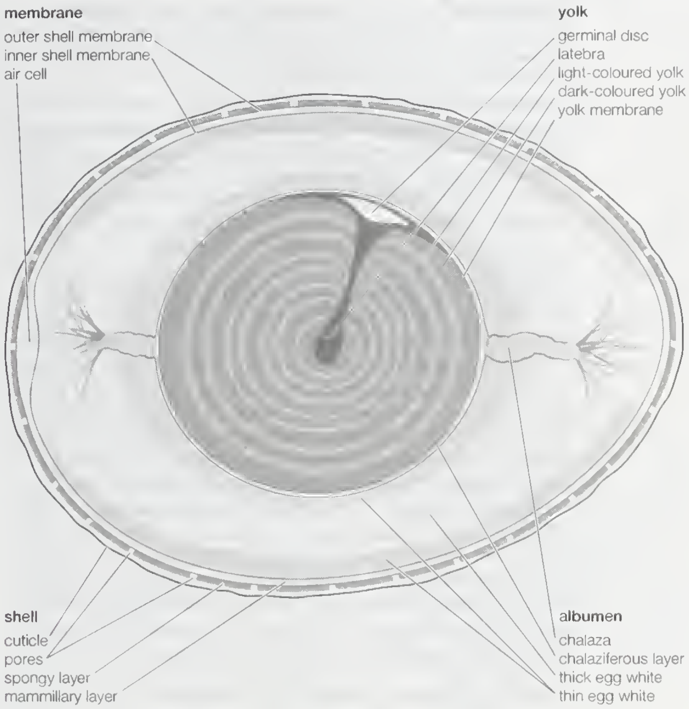


Figure 28: The structural components of an egg.
Encyclopædia Britannica, Inc.

The nutrient composition of chicken eggs is presented in Table 9. The whole egg is a source of high-quality protein (*i.e.,* proteins that contain all the amino acids needed in the human diet). In addition, it is an excellent source of all vitamins (except vitamin C) and contains many essential minerals, including phosphorus and zinc. All the fats, or lipids, as well as the cholesterol are found in the yolk. Yolk lipids are high in unsaturated fatty acids, with the ratio of unsaturated to saturated fatty acids commonly being 2 to 1. By influencing the diet of the hen, some processors are able to market shell eggs with yet a higher ratio of unsaturated to saturated fatty acids. Particular emphasis is being given to increasing the highly unsaturated long-chain omega-3 fatty acids by adding fish oil to the hen feed. Omega-3 fatty acids have been shown to play a role both in normal growth and development and in the prevention of many diseases.

The cholesterol content of a whole large egg is approximately 216 milligrams—a substantially lower figure than that reported before the late 1980s, when improved analytical techniques were instituted. Moreover, the egg industry has probably made some progress in lowering cholesterol content through genetic selection and improved diets.

Control of *Salmonella* contamination

Microbiology. More than 90 percent of all eggs are free of contamination at the time they are laid; contamination with *Salmonella* bacteria and with certain spoilage organisms occurs essentially afterward. Proper washing and sanitizing of eggs eliminates most *Salmonella* and spoilage organisms deposited on the shell. The organism *Salmonella enteritidis,* a common cause of gastroenteritis (a form of food poisoning), has been found to be trans-

Table 9: Nutrient Composition of Fresh Chicken Egg (per 100 grams)*

	energy (kcal)	water (g)	protein (g)	fat (g)	cholesterol (mg)	carbohydrate (g)	vitamin A (IU)	riboflavin (mg)	calcium (mg)	phosphorus (mg)
Whole egg	149	75.33	12.49	10.02	425	1.22	635	0.508	49	178
Yolk	358	48.81	16.76	30.87	1,281	1.78	1,945	0.639	137	488
White	50	87.81	10.52	0		1.03		0.452	6	13

*100 grams is approximately equal to two large whole eggs.
Source: *Composition of Foods,* Agriculture Handbook no. 8-1, U.S. Department of Agriculture.

ferred through the hen ovary in fewer than 1 percent of all eggs produced. Ovarian-transferred *S. enteritidis* can be controlled by thorough cooking of eggs (*i.e.*, until there are no runny whites or yolk).

Certain spoilage organisms (*e.g.*, *Alcaligenes*, *Proteus*, *Pseudomonas*, and some molds) may produce green, pink, black, colourless, and other rots in eggs after long periods of storage. However, since eggs move through market channels rapidly, the modern consumer seldom encounters spoiled eggs.

FRESH EGGS

Fresh eggs are gathered on automatic collection belts at the farm and stored in a cooler at about 7° C (45° F). The eggs are then delivered to a central processing plant, where they are washed, sanitized, and graded. Grading involves the sorting of eggs into size and quality categories using automated machines. Flash candling (passing the eggs over a strong light source) detects any abnormalities such as cracked eggs and eggs containing bloodspots or other defects. Higher-grade eggs have a thick, upstanding white, an oval yolk, and a clean, smooth, unbroken shell.

In the United States eggs are sized on the basis of a minimum weight per dozen in ounces. One dozen extra large eggs weigh 27 ounces (765 grams); large eggs, 24 ounces; medium eggs, 21 ounces. Weight standards in other countries vary, but most are measured in metric units. For example, eggs might be sold in cartons of 10 eggs each.

Most eggs sold in modern supermarkets are approximately four to five days old. If kept refrigerated by the consumer, they will maintain good quality and flavour for about four weeks.

EGG PRODUCTS

Although per capita consumption of fresh eggs has declined since 1950, the utilization of eggs in other food products has increased. As ingredients, egg products are tailored to suit the specific needs of the food processor. For example, the foaming properties of the white or yolk are important in bakery products; egg yolk serves as an emulsifier in mayonnaise and salad oils; and the addition of eggs to meats or other foods enhances their binding properties.

Egg products, in the form of liquid, dried, or frozen eggs, are used as ingredients in many kinds of food products. In addition, specialty egg products are sold as convenience foods directly to the consumer or to food-service establishments.

Liquid egg products. Refrigerated liquid egg products have become increasingly popular, especially in food-service establishments. Liquid egg products may be delivered in a variety of packages, including bulk tank trucks, smaller portable tanks or "totes," paper cartons, hermetically sealed polyethylene bags, lacquer-coated tins, and plastic pails. These products include liquid egg whites, liquid egg yolks, and various blends of the whites and yolks. Normally, liquid egg products are pasteurized at 60° C (140° F) for 3.5 minutes and have a shelf life of two to six days. Some liquid egg products are processed using ultra-pasteurization and aseptic packaging techniques to extend their shelf life to about six weeks.

Shelf life of liquid egg products

Dried egg products. Dried or dehydrated eggs are less expensive to ship, more convenient to use, and easier to store than fresh whole eggs. Spray dryers are used to produce a high-quality egg product with foaming and emulsification properties similar to those of fresh eggs. The dehydrated eggs are packed in containers ranging from small pouches to large drums, depending on their commercial application. Several types of dried egg products are produced for various applications in the food industry (*e.g.*, cake mixes, salad dressings, pasta). These products include dried egg white solids, instant egg white solids, stabilized (glucose removed) whole egg solids, and various blends of whole egg and yolk with sugar or corn syrup. Most dried egg products have a storage life of one year when refrigerated.

Frozen egg products. Frozen egg products are often preferred as ingredients in certain food products..Salt, sugar, or corn syrup is normally added to yolks or whole eggs prior to freezing in order to prevent gelation or thickening of the products. Egg whites freeze well without any additives. Egg products are frozen at −23° C (−9° F) and are packed in different-sized pouches and waxed or plastic cartons. Products include egg whites, egg yolks, salted yolks, sugared yolks, salted whole eggs, sugared whole eggs, and various yolk and white blends with or without added sugar or salt. At frozen temperatures they have a shelf life of about one year.

Specialty egg products. A number of specialty egg products are available to both individual consumers and institutions. Commercial salad bars utilize cryogenically frozen and diced hard-cooked eggs and pickled or plain hard-cooked eggs. Several frozen, precooked egg products are available in markets, including egg pizza, scrambled eggs, omelettes, French toast, breakfast sandwiches, crepes, and quiches. Several low-cholesterol or cholesterol-free egg substitutes have been developed by replacing the egg yolk with vegetable oils, emulsifiers, stabilizers, colour, vitamins, and minerals. Fat-free egg substitutes have also been developed for commercial use. (G.W.F.)

Dairy products

Milk has been used by humans since the beginning of recorded time to provide both fresh and storable nutritious foods. In some countries almost half the milk produced is consumed as fresh pasteurized whole, low-fat, or skim milk. However, most milk is manufactured into more stable dairy products of worldwide commerce, such as butter, cheese, dried milks, ice cream, and condensed milk.

Cow milk (bovine species) is by far the principal type used throughout the world. Other animals utilized for their milk production include buffalo (in India, China, Egypt, and the Philippines), goats (in the Mediterranean countries), reindeer (in northern Europe), and sheep (in southern Europe). This section focuses on the processing of cow milk and milk products unless otherwise noted. In general, the processing technology described for cow milk can be successfully applied to milk obtained from other species.

In the early 1800s the average dairy cow produced less than 1,500 litres of milk annually. With advances in animal nutrition and selective breeding, one cow now produces an average of 6,500 litres of milk a year, with some cows producing up to 10,000 litres. The Holstein-Friesian cow produces the greatest volume, but other breeds such as Ayrshire, Brown Swiss, Guernsey, and Jersey, while producing less milk, are known for supplying milk that contains higher fat, protein, and total solids.

Milk production

PROPERTIES OF MILK

Nutrient composition. Although milk is a liquid and most often considered a drink, it contains between 12 and 13 percent total solids and perhaps should be regarded as a food. In contrast, many "solid" foods, such as tomatoes, carrots, and lettuce, contain as little as 6 percent solids.

Many factors influence the composition of milk, including breed, genetic constitution of the individual cow, age of the cow, stage of lactation, interval between milkings, and certain disease conditions. Since the last milk drawn at each milking is richest in fat, the completeness of milking also influences a sample. In general, the type of feed only slightly affects the composition of milk, but feed of poor quality or insufficient quantity causes both a low yield and a low percentage of total solids. Current feeding programs utilize computer technology to achieve the greatest efficiency from each animal.

The composition of milk varies among mammals, primarily to meet growth rates of the individual species. The proteins contained within the mother's milk are the major components contributing to the growth rate of the young animals. Human milk is relatively low in both proteins and minerals compared with that of cows and goats.

Goat milk has about the same nutrient composition as cow milk, but it differs in several characteristics. Goat milk is completely white in colour because all the beta-carotene (ingested from feed) is converted to vitamin A. The fat globules are smaller and therefore remain suspended, so the cream does not rise and mechanical homogenization is unnecessary. Goat milk curd forms into small, light flakes and is more easily digested, much like the curd formed from human milk. It is often prescribed for persons who are allergic to the proteins in cow milk and for some patients afflicted with stomach ulcers.

Sheep milk is rich in nutrients, having 18 percent total solids (5.8 percent protein and 6.5 percent fat). Reindeer milk has the highest level of nutrients, with 36.7 percent total solids (10.3 percent protein and 22 percent fat). These high-fat, high-protein milks are excellent ingredients for cheese and other manufactured dairy products.

The major components of milk are water, fat, protein, carbohydrate (lactose), and minerals (ash). However, there are numerous other highly important micronutrients such as vitamins, essential amino acids, and trace minerals. Indeed, more than 250 chemical compounds have been identified in milk. Table 10 shows the composition of fresh fluid milk and other dairy products.

Milk fat

Fat. The fat in milk is secreted by specialized cells in the mammary glands of mammals. It is released as tiny fat globules or droplets, which are stabilized by a phospholipid and protein coat derived from the plasma membrane of the secreting cell. Milk fat is composed mainly of triglycerides—three fatty acid chains attached to a single molecule of glycerol. It contains 65 percent saturated, 32 percent monounsaturated, and 3 percent polyunsaturated fatty acids. The fat droplets carry most of the cholesterol and vitamin A. Therefore, skim milk, which has more than 99.5 percent of the milk fat removed, is significantly lower in cholesterol than whole milk (2 milligrams per 100 grams of milk, compared with 14 milligrams for whole milk) and must be fortified with vitamin A.

Protein. Milk contains a number of different types of proteins, depending on what is required for sustaining the young of the particular species. These proteins increase the nutritional value of milk and other dairy products and provide certain characteristics utilized for many of the processing methods. A major milk protein is casein, which actually exists as a multisubunit protein complex dispersed throughout the fluid phase of milk. Under certain conditions the casein complexes are disrupted, causing curdling of the milk. Curdling results in the separation of milk proteins into two distinct phases, a solid phase (the curds) and a liquid phase (the whey).

Lactose. Lactose is the principal carbohydrate found in milk. It is a disaccharide composed of one molecule each of the monosaccharides (simple sugars) glucose and galactose. Lactose is an important food source for several types of fermenting bacteria. The bacteria convert the lactose into lactic acid, and this process is the basis for several types of dairy products.

In the diet lactose is broken down into its component glucose and galactose subunits by the enzyme lactase. The glucose and galactose can then be absorbed from the digestive tract for use by the body. Individuals deficient in lactase cannot metabolize lactose, a condition called lactose intolerance. The unmetabolized lactose cannot be absorbed from the digestive tract and therefore builds up, leading to intestinal distress.

Vitamins and minerals. Milk is a good source of many vitamins. However, its vitamin C (ascorbic acid) content is easily destroyed by heating during pasteurization. Vitamin D is formed naturally in milk fat by ultraviolet irradiation but not in sufficient quantities to meet human nutritional needs. Beverage milk is commonly fortified with the fat-soluble vitamins A and D. In the United States the fortification of skim milk and low-fat milk with vitamin A (in water-soluble emulsified preparations) is required by law.

Milk also provides many of the B vitamins. It is an excellent source of riboflavin (B_2) and provides lesser amounts of thiamine (B_1) and niacin. Other B vitamins found in trace amounts are pantothenic acid, folic acid, biotin, pyridoxine (B_6), and vitamin B_{12}.

Milk is also rich in minerals and is an excellent source of calcium and phosphorus. It also contains trace amounts of potassium, chloride, sodium, magnesium, sulfur, copper, iodine, and iron. A lack of adequate iron is said to keep milk from being a complete food.

Enzymes

Physical and biochemical properties. Milk contains many natural enzymes, and other enzymes are produced in milk as a result of bacterial growth. Enzymes are biological catalysts capable of producing chemical changes in organic substances. Enzyme action in milk systems is extremely important for its effect on the flavour and body of different milk products. Lipases (fat-splitting enzymes), oxidases, proteases (protein-splitting enzymes), and amylases (starch-splitting enzymes) are among the more important enzymes that occur naturally in milk. These classes of enzymes are also produced in milk by microbiological action. In addition, the proteolytic enzyme (*i.e.*, protease) rennin, produced in calves' stomachs to coagulate milk protein and aid in nutrient absorption, is used to coagulate milk for manufacturing cheese.

The coagulation of milk is an irreversible change of its protein from a soluble or dispersed state to an agglomerated or precipitated condition. Its appearance may be associated with spoilage, but coagulation is a necessary step in many processing procedures. Milk may be coagulated by rennin or other enzymes, usually in conjunction

Table 10: Nutrient Composition of Dairy Products (per 100 grams)

dairy product	energy (kcal)	water (g)	protein (g)	fat (g)	carbohydrate (g)	cholesterol (mg)	vitamin A (IU)	riboflavin (mg)	calcium (mg)
Fresh fluid milk									
Whole	61	88	3.29	3.34	4.66	14	126	0.162	119
Low-fat*	50	89	3.33	1.92	4.80	8	205	0.165	122
Skim*	35	91	3.41	0.18	4.85	2	204	0.140	123
Evaporated milk	134	74	6.81	7.56	10.04	29	243	0.316	261
Evaporated skim milk*	78	79	7.55	0.20	11.35	4	392	0.309	290
Sweetened condensed milk	321	27	7.91	8.70	54.40	34	328	0.416	284
Nonfat dry milk*	358	4	35.10	0.72	52.19	18	2,370	1.744	1,231
Butter	717	16	0.85	81.11	0.06	219	3,058	0.034	24
Ice cream (vanilla)	201	61	3.50	11.00	23.60	44	409	0.240	128
Ice milk (vanilla)	139	68	3.80	4.30	22.70	14	165	0.265	139
Sherbet (orange)	138	66	1.10	2.00	30.40	5	76	0.068	54
Frozen yogurt (nonfat)	128	69	3.94	0.18	28.16	2	7	0.265	134
Buttermilk	40	90	3.31	0.88	4.79	4	33	0.154	116
Sour cream	214	71	3.16	20.96	4.27	44	790	0.149	116
Yogurt, plain (low-fat)	63	85	5.25	1.55	7.04	6	66	0.214	183
Yogurt, fruit (low-fat)	102	74	4.37	1.08	19.05	4	46	0.178	152
Cheese									
Blue	353	42	21.40	28.74	2.34	75	721	0.382	528
Brie	334	48	20.75	27.68	0.45	100	667	0.520	184
Cheddar	403	37	24.90	33.14	1.28	105	1,059	0.375	721
Cottage	103	79	12.49	4.51	2.68	15	163	0.163	60
Cream	349	54	7.55	34.87	2.66	110	1,427	0.197	80
Mozzarella**	280	49	27.47	17.12	3.14	54	628	0.343	731
Parmesan, grated	456	18	41.56	30.02	3.74	79	701	0.386	1,376
Emmentaler (Swiss)	376	37	28.43	27.45	3.38	92	845	0.365	961

*Fortified with vitamin A. **Low moisture, part skim.

Source: *Composition of Foods,* Agriculture Handbook no. 8-1, U.S. Department of Agriculture.

with heat. Left unrefrigerated, milk may naturally sour or coagulate by the action of lactic acid, which is produced by lactose-fermenting bacteria. This principle is utilized in the manufacture of cottage cheese. When milk is pasteurized and continuously refrigerated for two or three weeks, it may eventually coagulate or spoil owing to the action of psychrophilic or proteolytic organisms that are normally present or result from postpasteurization contamination.

© Larry Lefever/Grant Heilman Photography, Inc

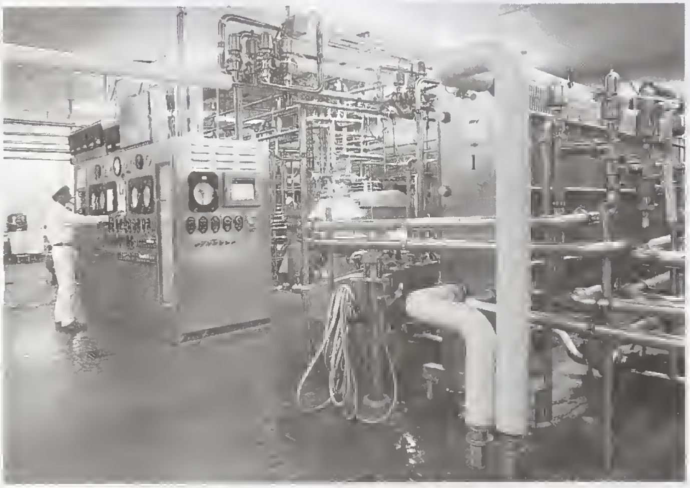


Figure 29: Equipment for the high-temperature short-time pasteurization of milk.

Milk fat is present in milk as an emulsion in a water phase. Finely dispersed fat globules in this emulsion are stabilized by a milk protein membrane, which permits the fat to clump and rise. The rising action is called creaming and is expected in all unhomogenized milk. In the United States, when paper cartons supplanted glass bottles, consumers stopped the practice of skimming cream from the top. Processors then introduced homogenization, a method of preventing gravity separation by forcing milk through very small openings under pressure, thus reducing fat globules to one-tenth their original size. Homogenization is practiced in many dairy processes in order to improve the physical properties of products (see below *Fresh fluid milk: Processing*).

Milk and other dairy products are very susceptible to developing off-flavours. Some flavours, given such names as "feed," "barny," or "unclean," are absorbed from the food ingested by the cow and from the odours in its surroundings. Others develop through microbial action due to growth of bacteria in large numbers. Chemical changes can also take place through enzyme action, contact with metals (such as copper), or exposure to sunlight or strong fluorescent light. Quality-control directors are constantly striving to avoid off-flavours in milk and other dairy foods.

FRESH FLUID MILK

Fresh fluid milk requires the highest-quality raw milk and is generally designated as Grade "A." This grade requires a higher level of sanitation and inspection on the farm than is necessary for "manufacturing grade" milk.

Quality concerns. Raw milk is a potentially dangerous food that must be processed and protected to assure its safety for humans. While most bovine diseases, such as brucellosis and tuberculosis, have been eliminated, many potential human pathogens inhabit the dairy farm environment. Therefore, it is essential that all milk be either pasteurized or (in the case of cheese) held for at least 60 days if made from raw milk. While milk from healthy cows is often totally bacteria-free, that condition quickly changes when milk is exposed to the farm environment.

Milk received at the processing plant is tested before being unloaded from either farm-based tank trucks or over-the-road tankers. The milk is checked for odour, appearance, proper temperature, acidity, bacteria, and the presence of drug residues. These tests take no longer than 10 to 15 minutes. If the tank load passes these tests, the milk is pumped into the plant's refrigerated storage tanks. The milk is then stored for the shortest possible time.

Processing. Essential steps in the processing of fluid milk into various dairy products are shown in Figure 30.

Pasteurization. Pasteurization is most important in all dairy processing. It is the biological safeguard which ensures that all potential pathogens are destroyed. Extensive studies have determined that heating milk to 63° C (145° F) for 30 minutes or 72° C (161° F) for 15 seconds kills the most resistant harmful bacteria. In actual practice these temperatures and times are exceeded, thereby not only ensuring safety but also extending shelf life.

High-temperature short-time pasteurization

Most milk today is pasteurized by the continuous high-temperature short-time (HTST) method (72° C or 161° F for 15 seconds or above). The HTST method is conducted in a series of stainless steel plates and tubes, with the hot pasteurized milk on one side of the plate being cooled by the incoming raw milk on the other side. (See Figure 29.) This "regeneration" can be more than 90 percent efficient and greatly reduces the cost of heating and cooling. There are many fail-safe controls on an approved pasteurizer system to ensure that all milk is completely heated for the full time and temperature requirement. If the monitoring instruments detect that something is wrong, an automatic flow diversion valve will prevent the milk from moving on to the next processing stage. Higher temperatures and sometimes longer holding times are required for the pasteurization of milk or cream with a high fat or sugar content.

Pasteurized milk is not sterile and is expected to contain small numbers of harmless bacteria. Therefore, the milk must be immediately cooled to below 4.4° C (40° F) and protected from any outside contamination. The shelf life for high-quality pasteurized milk is about 14 days when properly refrigerated.

Extended shelf life can be achieved through ultrapasteurization. In this case, milk is heated to 138° C (280° F) for two seconds and aseptically placed in sterile conventional milk containers. Ultrapasteurized milk and cream must be refrigerated and will last at least 45 days. This process does minimal damage to the flavour and extends the shelf life of slow-selling products such as cream, eggnog, and lactose-reduced milks.

Ultrahigh-temperature (UHT) pasteurization is the same heating process as ultrapasteurization (138° C or 280° F for two seconds), but the milk then goes into a more substantial container—either a sterile five-layer laminated "box" or a metal can. This milk can be stored without refrigeration and has a shelf life of six months to a year. Products handled in this manner do not taste as fresh, but they are useful as an emergency supply or when refrigeration is not available.

Centrifuge separators

Separation. Most modern plants use a separator to control the fat content of various products. A separator is a high-speed centrifuge that acts on the principle that cream or butterfat is lighter than other components in milk. (The specific gravity of skim milk is 1.0358, specific gravity of heavy cream 1.0083.) The heart of the separator is an airtight bowl with funnellike stainless steel disks. The bowl is spun at a high speed (about 6,000 revolutions per minute), producing centrifugal forces of 4,000 to 5,000 times the force of gravity. Centrifugation causes the skim, which is denser than cream, to collect at the outer wall of the bowl. The lighter part (cream) is forced to the centre and piped off for appropriate use.

An additional benefit of the separator is that it also acts as a clarifier. Particles even heavier than the skim, such as sediment, somatic cells, and some bacteria, are thrown to the outside and collected in pockets on the side of the separator. This material, known as "separator sludge," is discharged periodically and sometimes automatically when buildup is sensed.

Most separators are controlled by computers and can produce milk of almost any fat content. Current standards generally set whole milk at 3.25 percent fat, low-fat at 1 or 2 percent, and skim at less than 0.5 percent. (Most skim milk is actually less than 0.01 percent fat.)

Homogenization. Milk is homogenized to prevent fat globules from floating to the top and forming a cream layer or cream plug. Homogenizers are simply heavy-

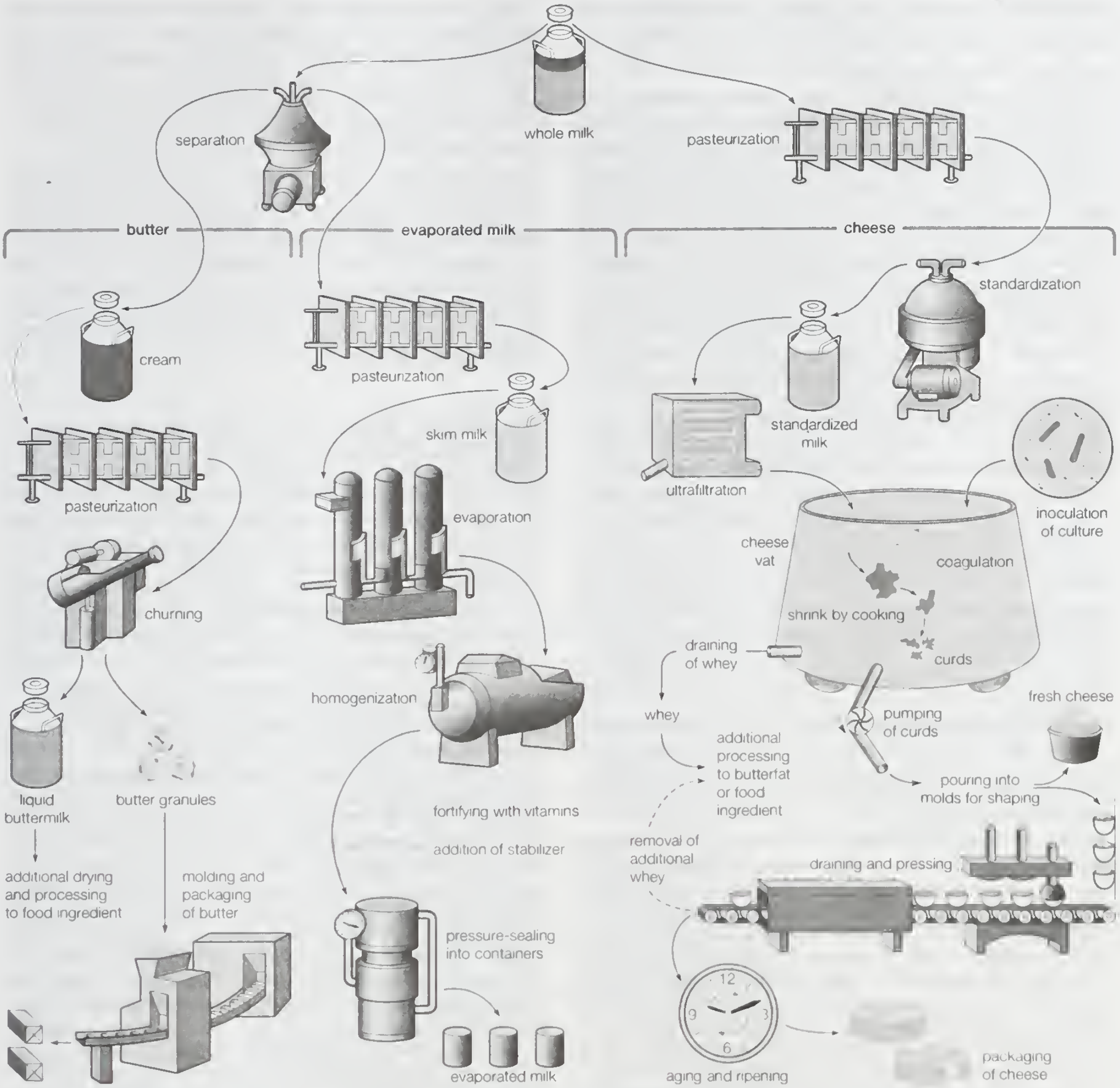


Figure 30: Essential steps in the processing of whole fresh milk into butter, evaporated milk, and cheese.
Encyclopædia Britannica, Inc.

duty, high-pressure pumps equipped with a special valve at the discharge end. They are designed to break up fat globules from their normal size of up to 18 micrometres to less than 2 micrometres in diameter (a micrometre is one-millionth of a metre). Hot milk (with the fat in liquid state) is pumped through the valve under high pressure, resulting in a uniform and stable distribution of fat throughout the milk.

Two-stage homogenization is sometimes practiced, during which the milk is forced through a second homogenizer valve or a breaker ring. The purpose is to break up fat clusters or clumps and thus produce a more uniform product with a slightly reduced viscosity.

Homogenization is considered successful when there is no visible separation of cream and the fat content in the top 100 millilitres of milk in a one-litre bottle does not differ by more than 10 percent from the bottom portion after standing 48 hours.

In addition to avoiding a cream layer, other benefits of homogenized milk include a whiter appearance, richer flavour, more uniform viscosity, better "whitening" in coffee, and softer curd tension (making the milk more digestible for humans). Homogenization is also essential for providing improved body and texture in ice cream, as well as numerous other products such as half-and-half, cream cheese, and evaporated milk.

Packaging. Until the mid 1880s milk was dipped from large cans into the consumer's own containers. The glass milk bottle was invented in 1884 and became the main container of retail distribution until World War II, when wax-coated paper containers were introduced. Plastic-coated paper followed and became the predominate container. Today more than 75 percent of retail sales are in translucent plastic jugs. Glass bottles make up less than 0.5 percent of the business and are used mostly at dairy stores and for home delivery.

Modern packaging machines are self-cleaning and provide an aseptic environment for milk packaging. Their improved design has allowed milk to remain fresh for at least 14 days and has made it possible for use with ultrapasteurizing equipment for extended shelf-life applications.

Specialty milks. Many specialty milks are now available (even in remote areas) as a result of the 45-day refrigerated shelf life of ultrapasteurized milk. One of the most useful products, lactose-reduced milk, is available in both nonfat and low-fat composition as well as in many flavoured versions. The lactose (milk sugar) is reduced by 70 to 100 percent, making it possible for lactose-intolerant individuals to enjoy the benefits of milk in their diets. Lactose reduction is accomplished by subjecting the appropriate milk to the action of the enzyme lactase in a refrigerated tank for approximately 24 hours. The enzyme breaks down the lactose to more readily digestible glucose and galactose. The reaction is halted when the lactose is consumed or when the milk is heat-treated. The resulting beverage is sweeter than regular milk but acceptable for most uses.

Lactose-reduced milk

Other specialty milks include calcium-fortified, special and seasonal flavours (*e.g.*, eggnog), and high-volume flavoured milk shakes (frequently served in schools).

CONDENSED AND DRIED MILK

Condensed and evaporated milk. Whole, low-fat, and skim milks, as well as whey and other dairy liquids, can be efficiently concentrated by the removal of water, using heat under vacuum. Since reducing atmospheric pressure lowers the temperature at which liquids boil, the water in milk is evaporated without imparting a cooked flavour. Water can also be removed by ultrafiltration and reverse osmosis, but this membrane technology is more expensive. Usually about 60 percent of the water is removed, which reduces storage space and shipping costs. Whole milk, when concentrated, usually contains 7.5 percent milk fat and 25.5 percent total milk solids. Skim milk can be condensed to approximately 20 to 40 percent solids, depending on the buyer's needs.

Condensed milk is often sold in refrigerated tank-truck loads to manufacturers of candy, bakery goods, ice cream, cheese, and other foods. When preserved by heat in individual cans, as shown in Figure 30, it is usually called "evaporated milk." In this process the concentrated milk is homogenized, fortified with vitamin D (A and D in evaporated skim milk), and sealed in a can sized for the consumer. A stabilizer, such as disodium phosphate or carrageenan, is also added to keep the product from separating during processing and storage. The sealed can is then sterilized at 118° C (244° F) for 15 minutes, cooled, and labeled. Evaporated milk keeps indefinitely, although staling and browning may occur after a year.

New ultrahigh-temperature (UHT) processing and aseptic filling of foil-lined cardboard or metal cans is also practiced. Although this process is more costly, the scorched flavour is not as pronounced as with conventionally processed evaporated milk.

Sweetened condensed milk

Sweetened condensed milk is also made by partially removing the water (as in evaporated milk) and adding sugar. The final product contains about 8.5 percent milk fat and at least 28 percent total milk solids. Sugar is added in sufficient amount to prevent bacterial action and subsequent spoilage. Usually, at least 60 percent sugar in the water phase is required to provide sufficient osmotic pressure for prevention of bacterial growth. Because sweetened condensed milk (or skim milk) is preserved by sugar, the milk merely needs to be pasteurized before being placed in a sanitary container (usually a metal can).

Dry milk products. Milk and by-products of milk production are often dried to reduce weight, to aid in shipping, to extend shelf life, and to provide a more useful form as an ingredient for other foods. In addition to skim and whole milk, a variety of useful dairy products are dried, including buttermilk, malted milk, instant breakfast, sweet cream, sour cream, butter powder, ice cream mix, cheese whey, coffee creamer, dehydrated cheese products, lactose, and caseinates. Many drying plants are built in conjunction with a butter-churning plant. These plants utilize the skim milk generated from the separated cream and the buttermilk produced from churning the butter. Most products are dried to less than 4 percent moisture to prevent bacterial growth and spoilage. However, products containing fat lose their freshness rather quickly owing to the oxidation of fatty acids, leading to rancidity.

Two types of dryers are used in the production of dried milk products—drum dryers and spray dryers. Each dryer has certain advantages.

Drum dryers. The simplest and least expensive is the drum, or roller, dryer. It consists of two large steel cylinders that turn toward each other and are heated from the inside by steam. The concentrated product is applied to the hot drum in a thin sheet that dries during less than one revolution and is scraped from the drum by a steel blade. The flakelike powder dissolves poorly in water but is often preferred in certain bakery products. Drum dryers are also used to manufacture animal feed where texture, flavour, and solubility are not a major consideration.

Spray dryers. Spray dryers are more commonly used since they do less heat damage and produce more soluble products. Concentrated liquid dairy product is sprayed in a finely atomized form into a stream of hot air. The air may be heated by steam-heated "radiators" or directly by sulfur-free natural gas. The drying chamber may be rectangular (the size of a living room), conical, or silo-shaped (up to five stories high). The powder passes from the drying chamber through a series of cyclone collectors and is usually placed in plastic-lined, heavy-duty paper bags.

Spray-dried milk is also difficult to reconstitute or mix with water. Therefore, a process called agglomeration was developed to "instantize" the powder, or make it more soluble. This process involves rewetting the fine, spray-dried powder with water to approximately 8 to 15 percent moisture and following up with a second drying cycle. The powder is now granular and dissolves very well in water. Virtually all retail packages of nonfat dry milk powder are instantized in this manner.

BUTTER

Composition. Butter is one of the most highly concentrated forms of fluid milk. Twenty litres of whole milk are needed to produce one kilogram of butter. This process leaves approximately 18 litres of skim milk and buttermilk, which at one time were disposed of as animal feed or waste. Today the skim portion has greatly increased in value and is fully utilized in other products.

Commercial butter is 80–82 percent milk fat, 16–17 percent water, and 1–2 percent milk solids other than fat (sometimes referred to as curd). It may contain salt, added directly to the butter in concentrations of 1 to 2 percent. Unsalted butter is often referred to as "sweet" butter. This should not be confused with "sweet cream" butter, which may or may not be salted. Reduced-fat, or "light," butter usually contains about 40 percent milk fat.

"Sweet" butter and "sweet cream" butter

Before World War II much of the butter produced in the United States was made from gathered cream. Farmers separated milk on the farm and shipped cans of cream to a butter factory, sometimes once or twice a week. The cream was often sour and needed to be neutralized (with sodium hydroxide) before churning. When transportation and the value of the skim portion improved, whole milk was shipped to the creamery, providing a supply of "sweet cream" (*i.e.*, cream that had not soured) for butter making. With these improvements came the advent of higher-quality butter and the demise of naturally soured buttermilk. Virtually all butter in the United States today is sweet cream butter. A notable exception is butter made from whey cream salvaged in the cheese-making process. The quality of fresh whey cream butter is indistinguishable from sweet cream butter.

Production. As shown in Figure 30, butter is produced when the cream emulsion in unhomogenized milk is destabilized by agitation, or churning. Breaking the emulsion produces butterfat granules the size of rice grains. The granules mat together and separate from the water phase or serum, which is known as buttermilk. (This milky liquid is drained away and is either concentrated or dried, later to become an ingredient in ice cream, candy, or other foods.) The butterfat is then washed with clean water and "worked" (kneaded) until more buttermilk separates and is removed. Ultimately, only about 16 percent of the water and milk solids present in the original milk remains trapped in the butter.

The churning process can take 40 to 60 minutes to complete in a traditional churn, but butter is more commonly made by high-speed continuous "churns" in factories. Although the basic principle is the same, in the continuous churn cream is pumped into a cylinder and mixed by high-speed blades, forming butter granules in seconds. The butter granules are forced through perforated plates while the buttermilk is drained from the system. A salt solution may be added if salted butter is desired. The butter is then worked in a twin screw extruder and emerges ready to be packaged.

Quality concerns. The quality of butter is based on its body, texture, flavour, and appearance. In the United States the Department of Agriculture (USDA) assigns quality grades to butter based on its score on a standard quality point scale. Grade AA is the highest possible grade; Grade AA butter must achieve a numerical score of 93

Grades of butter

out of 100 points based on its aroma, flavour, and texture. Salt (if present) must be completely dissolved and thoroughly distributed. Grade A butter is almost as good, with a score of 92 out of 100 points. Grade B butter is based on a score of 90 points, and it usually is used only for cooking or manufacturing. The flavour of Grade B is not as fresh and sweet, and its body may be crumbly, watery, or sticky.

Additions and treatment. The addition of salt to butter contributes to its flavour and also acts as a preservative. Added in concentrations of approximately 2 percent, all the salt goes into solution in the water phase. Since the water content of butter is less than 16 percent of the total volume, each water droplet can contain more than 10 percent salt. Such a concentration in the water phase limits bacterial growth overall, since the fat portion of butter is generally safe from microbial degradation.

Butter may contain added colouring. Butter from cows that are eating dry, stored feed during the winter may not contain enough beta-carotene for proper colouring, as it does when cows are pasture-fed. In such cases small amounts of a yellow vegetable colouring from the seed of the annatto tree may be added to enhance the colour.

Because butter is so firm when first removed from the refrigerator, it is sometimes whipped to improve spreadability. Generally, volume is increased by 50 percent by whipping in air—or, better still, nitrogen or an inert gas in order to prevent oxidation of the fat. Whipped butter, both salted and sweet, is sold in small plastic-coated tubs.

ICE CREAM AND OTHER FROZEN DESSERTS

Ice cream evolved from flavoured ices that were popular with the Roman nobility in the 4th century BC. The emperor Nero is known to have imported snow from the mountains and topped it with fruit juices and honey. In the 13th century Marco Polo was reported to have returned from China with recipes for making water and milk ices.

The discovery that salt would lower the freezing point of cracked ice led to the first practical method of making ice cream. Making ice cream in the home was greatly simplified by the invention of the wooden bucket freezer with rotary paddles. In 1851 the first wholesale ice cream was manufactured in Baltimore. With the development of mechanical refrigeration, widespread distribution of ice cream became possible. Ice cream parlours and drugstore soda counters flourished. With refrigerator-freezers now a standard domestic appliance, more than half of all frozen desserts are consumed at home.

Composition of frozen desserts. Standards for ice cream and most frozen desserts are closely regulated. In the United States, for example, ice cream must contain at least 10 percent fat and 20 percent total milk solids. In freezing, the volume may be doubled by the inclusion of air (known as overrun), but the increase in volume is limited to 100 percent by the requirement that the finished product weigh at least 4.5 pounds per gallon. Total food solids must weigh 1.6 pounds per gallon, thus limiting the water content. Regulations also require all ingredients to be listed, with some additives (such as stabilizers) limited to very small amounts.

The principal frozen desserts are ice cream, frozen custard, ice milk, frozen yogurt, sherbet, and water ices. Ice cream has the highest fat content, ranging from 10 to 20 percent. Frozen custard, or French ice cream, is basically the same formula as ice cream but contains added eggs or egg solids (usually 1.4 percent by weight). Ice milk may be more commonly called "light" or "reduced-fat" ice cream. It contains between 2 and 7 percent fat and at least 11 percent total milk solids. Frozen yogurt is a cultured frozen product containing the same ingredients as ice cream. It must contain at least 3.25 percent milk fat and 8.25 percent milk solids other than fat and must weigh at least five pounds per gallon. Low-fat frozen yogurt contains between 0.5 and 2 percent milk fat. Nonfat frozen yogurt is limited to less than 0.5 percent milk fat. Frozen yogurts should always contain live cultures of bacteria (see below *Cultured dairy foods: Yogurt*). The target overrun for ice cream, ice milk, and frozen yogurt is 65 to 100 percent. Premium ice creams may be as low as 20 percent overrun, while soft ice creams are in the 30 to 50 percent range.

Frozen yogurt

Sherbets contain relatively small quantities of milk products. Most standards require between 1 and 2 percent milk fat and between 2 and 5 percent total milk solids. Sherbet contains considerably more sugar and less air than ice cream (the target overrun is 30 to 40 percent), and therefore it is heavier and often contains more calories per serving. Water ices are similar to sherbet, but they contain no milk solids and have a target overrun of 20 to 30 percent.

Imitation ice cream, known as mellorine, is made in some parts of the United States and other countries. It is made with less expensive vegetable oils instead of butterfat but utilizes dairy ingredients for the milk protein part. Mellorines are intended to compete with ice cream in places where butterfat prices are high.

Ice cream manufacture. The essential ingredients in ice cream are milk, cream, sugar, flavouring, and stabilizer. Cheaper ingredients such as dry whey, corn syrup, and artificial flavourings may be substituted to create a lower-cost product.

The first step in ice cream making is formulating a suitable mix. The mix is composed of a combination of dairy ingredients, such as fresh milk and cream, frozen cream, condensed or dried skim, buttermilk, dairy whey, or whey protein concentrate. Sugars may include sucrose, corn syrup, honey, and other syrups. Stabilizers and emulsifiers are added in small amounts to help prevent formation of ice crystals, particularly during temperature fluctuations in storage.

The ice cream mix is pasteurized at no less than 79° C (175° F) for 25 seconds. The heated mix is typically homogenized in order to assure a smoother body and texture. Homogenizing also prevents churning (*i.e.*, separating out of fat granules) of the mix in the freezer and increases the viscosity. (Since smaller fat globules have more surface area, the associated milk protein can hydrate more water and produce a more viscous fluid.)

After homogenization, the hot mix is quickly cooled to 4.4° C (40° F). The mix must age at this temperature for at least four hours to allow the fat to solidify and fat globules to clump. This aging process results in quicker freezing and a smoother product.

Freezing

The next step, freezing the mix, is accomplished by one of two methods: continuous freezing, which uses a steady flow of mix, or batch freezing, which makes a single quantity at a time. For both methods, the objective is to freeze the product partially and, at the same time, incorporate air. The freezing process is carried out in a cylindrical barrel that is cooled by a refrigerant, either ammonia or Freon (trademark). The barrel is equipped with stainless steel blades, called dasher blades, which scrape the frozen mixture from the sides of the freezing cylinder and incorporate or whip air into the product. The amount of air incorporated during freezing is controlled by a pump or the dasher speed. Depending on individual conditions, freezing can be instantaneous in the continuous freezer or require approximately 10 minutes in the batch freezer.

Semifrozen ice cream leaves the freezer at a temperature between −9° and −5° C (16° and 23° F). It is placed in a suitable container and conveyed to a blast freezer, where temperatures are in the range of −29° to −34° C (−20° to −30° F). While in this room, the ice cream continues to freeze without agitation. Rapid freezing at this stage prevents the formation of large ice crystals and favours a smooth body and texture. The length of time in the hardening room depends on the size of the package, but usually 6 to 12 hours are required to bring the internal ice cream temperature to −18° C (0° F) or below. In larger manufacturing plants, final freezing takes place in a hardening tunnel, where packages are automatically conveyed on a continuous belt to the final storage area.

Much of the appeal of ice cream comes from the variety of standard and holiday flavours available throughout the year. Most ice cream manufacturers make a standard mix consisting of milk, cream, sugars, and stabilizers, to which flavours may be added just prior to freezing. High-volume flavours such as vanilla, chocolate, and strawberry may be blended in their own large batch tanks. For flavours with large particles, such as fruit, nuts, cookies, or candy parts, a "feeder" on the outlet of the freezer is used to inject the material. Ingredients such as fruits and nuts are carefully selected and specially treated to avoid introducing unwanted bacteria into the already pasteurized mix.

Ice cream and other frozen desserts require no preservatives and have long shelf lives if they are kept below −23° C (−10° F) and are protected from temperature fluctuations. Airtight packaging materials have made it possible to consider frozen storage of six months or longer without loss of flavour or body and texture. When ice cream is finally dipped, composition and overrun will determine ideal scooping temperature. This can vary from −16° to −9° C (3° to 15° F), with lower temperatures resulting in less dipping loss but more effort on the part of the server.

Ice cream can also be freeze-dried by the removal of 99 percent of the water. Freeze-drying eliminates the need for refrigeration and provides a high-energy food for hikers and campers and a "filling" centre for candy and other confections.

CULTURED DAIRY FOODS

With the development of microbiological and nutritional sciences in the late 19th century came the technology necessary to produce cultured dairy products on an industrial or commercial basis. Fermented milks had been made since early times, when warm raw milk from cows, sheep, goats, camels, or horses was naturally preserved by common strains of *Streptococcus* and *Lactobacillus* bacteria. (The "cultures" were obtained by including a small portion from the previous batch.) These harmless lactic acid producers were effective in suppressing spoilage and pathogenic organisms, making it possible to preserve fresh milk for several days or weeks without refrigeration. Cultured products eventually became ethnic favourites and were introduced around the world as people migrated.

Conversion of lactose by bacteria

Central to the production of cultured milk is the initial fermentation process, which involves the partial conversion of lactose (milk sugar) to lactic acid. Lactose conversion is accomplished by lactic-acid–producing *Streptococcus* and *Lactobacillus* bacteria. At temperatures of approximately 32° C (90° F), these bacteria reproduce very rapidly, perhaps doubling their population every 20 minutes. Many minute by-products that result from their metabolic processes assist in further ripening and flavouring of the cultured product. Subsequent or secondary fermentations can result in the production of other compounds, such as diacetyl (a flavour compound found in buttermilk) and alcohol (from yeasts in kefir), as well as butyric acid (which causes bitter or rancid flavours).

Cultured buttermilk, sour cream, and yogurt are among the most common fermented dairy products in the Western world. Other, lesser-known products include kefir, koumiss, acidophilus milk, and new yogurts containing *Bifidobacteria*. Cultured dairy foods provide numerous potential health benefits to the human diet. These foods are excellent sources of calcium and protein. In addition, they may help to establish and maintain beneficial intestinal bacterial flora and reduce lactose intolerance.

Buttermilk. Because of its name, most people assume buttermilk is high in fat. Actually, the name refers to the fact that buttermilk was once the watery end-product of butter making (shown in Figure 30). Modern buttermilk is made from low-fat or skim milk and has less than 2 percent fat and sometimes none. Its correct name in many jurisdictions is "cultured low-fat milk" or "cultured nonfat milk."

The starting ingredient for buttermilk is skim or low-fat milk. The milk is pasteurized at 82° to 88° C (180° to 190° F) for 30 minutes, or at 90° C (195° F) for two to three minutes. This heating process is done to destroy all naturally occurring bacteria and to denature the protein in order to minimize wheying off (separation of liquid from solids).

The milk is then cooled to 22° C (72° F), and starter cultures of desirable bacteria, such as *Streptococcus lactis, S. cremoris, Leuconostoc citrovorum,* and *L. dextranicum,* are added to develop buttermilk's acidity and unique flavour. These organisms may be used singly or in combination to obtain the desired flavour.

The ripening process takes about 12 to 14 hours (overnight). At the correct stage of acid and flavour, the product is gently stirred to break the curd, and it is cooled to 7.2° C (45° F) in order to halt fermentation. It is then packaged and refrigerated.

Sour cream. Sour cream is made according to the same temperature and culture methods as used for buttermilk. The main difference is the starting material—sour cream starts with light 18 percent cream.

Yogurt. Yogurt is made in a similar fashion to buttermilk and sour cream, but it requires different bacteria and temperatures. Whole, low-fat, or skim milk is fortified with nonfat dry milk or fresh condensed skim milk, in order to raise the total solids to 14 to 16 percent. The mixture is heat-treated as for buttermilk and then cooled to 45.6° to 46.7° C (114° to 116° F). At this point a culture of equal parts *Lactobacillus bulgaricus* and *Streptococcus thermophilus* is added to the warm milk, followed by one of two processing methods. For set, or sundae-style, yogurt (fruit on the bottom), the cultured mixture is poured into cups containing the fruit, held in a warm room until the milk coagulates (usually about four hours), and then moved to a refrigerated room. For blended (Swiss- or French-style) yogurt, the milk is allowed to incubate in large heated tanks. After coagulation occurs, the mixture is cooled, fruit or other flavours are added, and the product is placed in containers and immediately made ready for sale.

Set yogurt and blended yogurt

Many yogurt manufacturers have added *Lactobacillus acidophilus* to their bacterial cultures. *L. acidophilus* has possible health benefits in easing yeast infections and restoring normal bacterial balance to the intestinal tract of humans after antibiotic treatment.

CHEESE

Primitive forms of cheese have been made since humans started domesticating animals. No one knows exactly who made the first cheese, but, according to one ancient legend, it was made accidentally by an Arabian merchant crossing the desert. The merchant put his drinking milk in a bag made from a sheep's stomach. The natural rennin in the lining of the pouch, along with the heat from the sun, caused the milk to coagulate and then separate into curds and whey. At nightfall, the whey satisfied the man's thirst, and the curd (cheese) had a delightful flavour and satisfied his hunger.

From its birthplace in the Middle East, cheese making spread as far as England with the expansion of the Roman Empire. During the Middle Ages, monks and merchants of Europe made cheese an established food of that area. In 1620, cheese and cows were part of the ship's stores carried to North America by the Pilgrims on the *Mayflower*. Until the middle of the 19th century, cheese was a local farm product. Few, if any, distinct varieties of cheese were developed deliberately. Rather, cheese makers in each locality made a cheese that, when ripened under specific conditions of air temperature and humidity, mold, and milk source, acquired certain characteristics of its own. Different varieties appeared largely as a result of accidental changes or modifications in one or more steps of the cheese-making process. Because there was little understanding of the bacteriology and chemistry involved, these changes were little understood and difficult to duplicate. Cheese making was an art, and the process was a closely guarded secret that was passed down from one generation to the next.

With increasing scientific knowledge came a greater understanding of the bacteriological and chemical changes that are necessary to produce many types of cheese. Thus, it has become possible to control more precisely each step in the cheese-making process and to manufacture a more uniform product. Cheese making is now a science as well as an art.

Fundamentals of cheese making. The cheese-making process consists of removing a major part of the water contained in fresh fluid milk while retaining most of the solids. Since storage life increases as water content decreases, cheese making can also be considered a form of food preservation through the process of milk fermentation. The fermentation of milk into finished cheese requires several essential steps: preparing and inoculating the milk with lactic-acid–producing bacteria, curdling the milk, cutting the curd, shrinking the curd (by cooking), draining or dipping the whey, salting, pressing, and ripening. These steps, which are illustrated in Figure 30, begin with four basic ingredients: milk, microorganisms, rennet, and salt.

Inoculation and curdling. Milk for cheese making must be of the highest quality. Because the natural microflora present in milk frequently include undesirable types called psychrophiles, good farm sanitation and pasteurization or partial heat treatment are important to the cheese-making process. In addition, the milk must be free of substances that may inhibit the growth of acid-forming bacteria (*e.g.,* antibiotics and sanitizing agents). Milk is often pasteurized to destroy pathogenic microorganisms and to eliminate spoilage and defects induced by bacteria. However, since pasteurization destroys the natural enzymes found in milk, cheese produced from pasteurized milk ripens less rapidly and less extensively than most cheese made from raw or lightly heat-treated milk.

During pasteurization, the milk may be passed through a standardizing separator to adjust the fat-to-protein ratio of the milk. In some cases the cheese yield is improved by concentrating protein in a process known as ultrafiltration. The milk is then inoculated with fermenting microorganisms and rennet, which promote curdling.

Rennet

The fermenting microorganisms carry out the anaerobic conversion of lactose to lactic acid. The type of organisms used depends on the variety of cheese and on the production process. Rennet is an enzymatic preparation that is usually obtained from the fourth stomach of calves. It contains a number of proteolytic (protein-degrading) enzymes, including rennin and pepsin. Some cheeses, such as cottage cheese and cream cheese, are produced by acid coagulation alone. In the presence of lactic acid, rennet, or both, the milk protein casein clumps together and precipitates out of solution; this is the process known as curdling, or coagulation. Coagulated casein assumes a solid or gellike structure (the curd), which traps most of the fat, bacteria, calcium, phosphate, and other particulates. The remaining liquid (the whey) contains water, proteins resistant to acidic and enzymatic denaturation (*e.g.,* antibodies), carbohydrates (lactose), and minerals.

Lactic acid produced by the starter culture organisms has several functions. It promotes curd formation by rennet (the activity of rennet requires an acidic pH), causes the curd to shrink, enhances whey drainage (syneresis), and helps prevent the growth of undesirable microorganisms during cheese making and ripening. In addition, acid affects the elasticity of the finished curd and promotes fusion of the curd into a solid mass. Enzymes released by the bacterial cells also influence flavour development during ripening.

Salt is usually added to the curd. In addition to enhancing flavour, it helps to withdraw the whey from the curd and inhibits the growth of undesirable microorganisms.

Cutting and shrinking. After the curd is formed, it is cut with fine wire "knives" into small cubes approximately one centimetre (one-half inch) square. The curd is then gently heated, causing it to shrink. The degree of shrinkage determines the moisture content and the final consistency of the cheese. Whey is removed by draining or dipping. The whey may be further processed to make whey cheeses (*e.g.,* ricotta) or beverages, or it may be dried in order to preserve it as a food ingredient.

Ripening. Most cheese is ripened for varying amounts of time in order to bring about the chemical changes necessary for transforming fresh curd into a distinctive aged cheese. These changes are catalyzed by enzymes from three main sources: rennet or other enzyme preparations of animal or vegetable origin added during coagulation, microorganisms that grow within the cheese or on its surface, and the cheese milk itself. The ripening time may be as short as one month, as for Brie, or a year or more, as in the case of sharp cheddar.

The ripening of cheese is influenced by the interaction of bacteria, enzymes, and physical conditions in the curing room. The speed of the reactions is determined by temperature and humidity conditions in the room as well as by the moisture content of the cheese. In most cheeses lactose continues to be fermented to lactic acid and lactates, or it is hydrolyzed to form other sugars. As a result, aged cheeses such as Emmentaler and cheddar have no residual lactose.

In a similar manner, proteins and lipids (fats) are broken down during ripening. The degree of protein decomposition, or proteolysis, affects both the flavour and the consistency of the final cheese. It is especially apparent in Limburger and some blue-mold ripened cheeses. Surface-mold ripened cheeses, such as Brie, rely on enzymes produced by the white *Penicillium camemberti* mold to break down proteins from the outside. When lipids are broken down (as in Parmesan and Romano cheeses), the process is called lipolysis.

Secondary fermentation

The eyes, or holes, typical of Swiss-type cheeses such as Emmentaler and Gruyère come from a secondary fermentation that takes place when, after two weeks, the cheeses are moved from refrigerated curing to a warmer room, where temperatures are in the range of 20° to 24° C (68° to 75° F). At this stage, residual lactates provide a suitable medium for propionic acid bacteria (*Propionibacterium shermanii*) to grow and generate carbon dioxide gas. Eye formation takes three to six weeks. Warm-room curing is stopped when the wheels develop a rounded surface and the echo of holes can be heard when the cheese is thumped. The cheese is then moved back to a cold room, where it is aged at about 7° C (45° F) for 4 to 12 months in order to develop its typical sweet, nutty flavour.

The unique ripening of blue-veined cheeses comes from the mold spores *Penicillium roqueforti* or *P. glaucum,* which are added to the milk or to the curds before pressing and are activated by air. Air is introduced by "needling" the cheese with a device that punches about 50 small holes into the top. These air passages allow mold spores to grow vegetative cells and spread their greenish blue mycelia, or threadlike structures, through the cheese. *Penicillium* molds are also rich in proteolytic and lipolytic enzymes, so that during ripening a variety of trace compounds also are produced, such as free amines, amino acids, carbonyls, and fatty acids—all of which ultimately affect the flavour and texture of the cheese.

Surface-ripened cheeses such as Gruyère, brick, Port Salut, and Limburger derive their flavour from both internal ripening and the surface environment. For instance,

Table 11: Varieties of Cheese, Classified by Hardness and Ripening Method

	ripening method	cheese variety
Very hard	bacteria/enzymes	Asiago, Parmesan, Romano, Sapsago, Sonoma Dry Jack
Hard	bacteria/enzymes	Cantal, cheddar, Colby
	eye-producing bacteria/enzymes	Emmentaler (Swiss), Gruyère, Fontina, Jarlsberg
Semihard/semisoft	bacteria/enzymes	brick, Edam, Gouda, Monterey Jack, mozzarella, Munster, provolone
	bacteria/enzymes and surface microorganisms	Bel Paese, brick, Limburger, Port Salut, Trappist
	bacteria/enzymes and blue mold	blue, Gorgonzola, Roquefort, Stilton
Soft	bacteria/enzymes and surface microorganisms	Brie, Camembert, Neufchâtel (France), Pont l'Évêque
	unripened	baker's, cottage, cream, feta, Neufchâtel (United States), pot

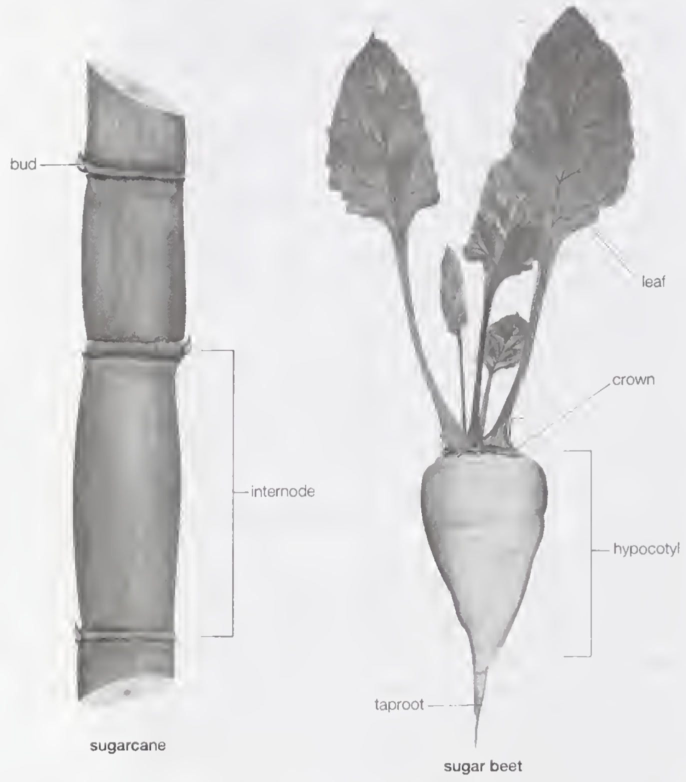


Figure 31: Structures of the sugarcane (left) and sugar beet (right).
Encyclopædia Britannica, Inc.

the high-moisture wiping of the surface of Gruyère gives that cheese a fuller flavour than its Emmentaler counterpart. Specific organisms, such as *Brevibacterium linens,* in Limburger cheese result in a reddish brown surface growth and the breakdown of protein to amino nitrogen. The resulting odour is offensive to some, but the flavour and texture of the cheese are pleasing to many.

Not all cheeses are ripened. Cottage, cream, ricotta, and most mozzarella cheeses are ready for sale as soon as they are made. All these cheeses have sweet, delicate flavours and often are combined with other foods.

Varieties of cheese. As a result of the many combinations of milks, cultures, enzymes, molds, and technical processes, literally hundreds of varieties of cheese are made throughout the world. The different types of cheese can be classified in many ways; the most effective is probably according to hardness or ripening method. Table 11 groups several varieties of cheese based on these criteria.

In recent years different types of cheese have been combined in order to increase variety and consumer interest. For example, soft and mildly flavoured Brie is combined with a more pungent semisoft cheese such as blue or Gorgonzola. The resulting "Blue-Brie" has a bloomy white edible rind, while its interior is marbled with blue *Penicillium roqueforti* mold. The cheese is marketed under various names such as Bavarian Blue, Cambazola, Lymeswold, and Saga Blue. Another combination cheese is Norwegian Jarlsberg. This cheese results from a marriage of the cultures and manufacturing procedures for Dutch Gouda and Swiss Emmentaler.

Pasteurized process cheese. Some natural cheese is made into process cheese, a product in which complete ripening is halted by heat. The resulting product has an indefinite shelf life. Most process cheese is used in food service outlets and other applications where convenient, uniform melting is required.

Process cheeses and process cheese foods

Pasteurized process cheese is made by grinding and mixing natural cheese with other ingredients, such as water, emulsifying agents, colouring, fruits, vegetables, or meat. The mixture is then heated to temperatures of 165° F (74° C) and stirred into a homogeneous, plastic mass. Process cheese foods, spreads, and products differ from process cheese in that they may contain other ingredients, such as nonfat dry milk, cheese whey, and whey protein concentrates, as well as additional amounts of water.

American cheddar is processed most frequently. However, other cheeses such as washed-curd, Colby, Swiss, Gruyère, and Limburger are similarly processed. In a slight variation, cold pack or club cheese is made by grinding and mixing together one or more varieties of cheese without heat. This cheese food may contain added flavours or ingredients. (D.K.B.)

Sugar

Sugar is the common name for sucrose, a crystalline tabletop and industrial sweetener used in foods and beverages. As a chemical term, "sugar" usually refers to all carbohydrates of the general formula $C_n(H_2O)_n$. Sucrose is a disaccharide, or double sugar, being composed of one molecule of glucose linked to one molecule of fructose. Because one molecule of water (H_2O) is lost in the condensation reaction linking glucose to fructose, sucrose is represented by the formula $C_{12}H_{22}O_{11}$ (following the general formula $C_n[H_2O]_{n-1}$).

Sucrose is found in almost all plants, but it occurs at concentrations high enough for economic recovery only in sugarcane (*Saccharum officinarum*) and sugar beets (*Beta vulgaris*). The former is a giant grass growing in tropical and subtropical areas; the latter is a root crop growing in temperate zones (see Figure 31). Sugarcane ranges from 7 to 18 percent sugar by weight, while sugar beets are from 8 to 22 percent sugar by weight. Sucrose from either source (or from two relatively minor sources, the sugar maple tree and the date palm) is the same molecule, yielding 3.94 calories per gram as do all carbohydrates. Differences in sugar products come from other components isolated with sucrose.

Sugar content of cane and beet

The first cultivated sugar crop was sugarcane, developed from wild varieties in the East Indies—probably New Guinea. The sugar beet was developed as a crop in Europe in the 19th century during the Napoleonic Wars, when France sought an alternate homegrown source of sugar in order to save its ships from running blockades to sugarcane sources in the Caribbean. Sugarcane, once harvested, cannot be stored because of sucrose decomposition. For this reason, cane sugar is generally produced in two stages, manufacture of raw sugar taking place in the cane-growing areas and refining into food products occurring in the sugar-consuming countries. Sugar beets, on the other hand, can be stored and are therefore generally processed in one stage into white sugar.

CANE SUGAR

Cane harvesting and delivery. Sugarcane is generally harvested in the cooler months of the year, although it is harvested year-round in Cuba, the Philippines, Colombia, and other prime areas. As much as two-thirds of the world's cane crop is harvested by hand, using long machetes. Since the 1940s, however, mechanical harvesting has increased. Before or after harvest, the cane is burned in order to drive out rodents and snakes and to burn off leaves and trash that dull knife blades, but environmental considerations are leading to the harvesting of whole unburned cane in several areas.

Harvested cane is transported to the factory by many means, ranging from manual haulage to oxcarts, trucks, railway cars, or barges. The usual economic distance between field and factory is 25 kilometres (15 miles). Minimizing the time between cutting and processing reduces the amount of cane deterioration and encourages a higher sugar yield.

Upon arrival at the factory gate, cane is weighed and sampled for analysis (if factors other than weight are used for payment). Cane is stored in as small amounts and for as short a time as possible in the mill yard. Factories run around the clock, stopping in some areas for only one or two days per month for cleaning. Although payment is usually based on weight and sucrose content, quality factors such as moisture, trash, and fibre content also are included. Payment is generally split, with 60 to 65 percent going to the grower and 35 to 40 percent going to the processor.

Raw sugar manufacture. Sugarcane processing, outlined in Figure 32, is practiced in many variations, but the essential process consists of the following steps: extraction of the cane juice by milling or diffusion, clarification of the juice, concentration of the juice to syrup by evaporation, crystallization of sugar from the syrup, and separation and drying of the crystals.

Juice extraction. After weighing, sugarcane is loaded by hand or crane onto a moving table. The table carries the cane into one or two sets of revolving knives, which chop the cane into chips in order to expose the tissue and open the cell structure, thus readying the material for efficient extraction of the juice. Frequently, knives are followed by a shredder, which breaks the chips into shreds for finer cane preparation. The chipped (and shredded) cane then goes through the crusher, a set of roller mills in which the cane cells are crushed and juice extracted. As the crushed cane proceeds through a series of up to eight four-roll mills, it is forced against a countercurrent of water known as water of maceration or imbibition. Streams of juice extracted from the cane, mixed with maceration water from all mills, are combined into a mixed juice called dilute juice. Juice from the last mill in the series (which does not receive a current of maceration water) is called residual juice.

Dilute juice

The alternative to extraction by milling is extraction by diffusion. In this process, cane prepared by rotating knives and a shredder is moved through a multicell, countercurrent diffuser. Extraction of sugar is higher by diffusion (an average rate of 93 percent, compared with 85–90 percent by milling), but extraction of nonsugars is also higher. Diffusion, therefore, is most used where cane quality is highest—*e.g.,* in South Africa, Australia, and Hawaii. Occasionally a smaller "bagasse diffuser" is used in order to increase extraction from partially milled cane after two or three mills. (Residual cane fibre, after juice is removed, is called bagasse.)

Disposal of the large amounts of water used by diffusers is a costly environmental problem, as cane factories that practice diffusion must operate their own primary, secondary, and tertiary water-treatment systems.

Clarification. Mixed juice from the extraction mills or diffuser is purified by addition of heat, lime, and flocculation aids. The lime is a suspension of calcium hydroxide,

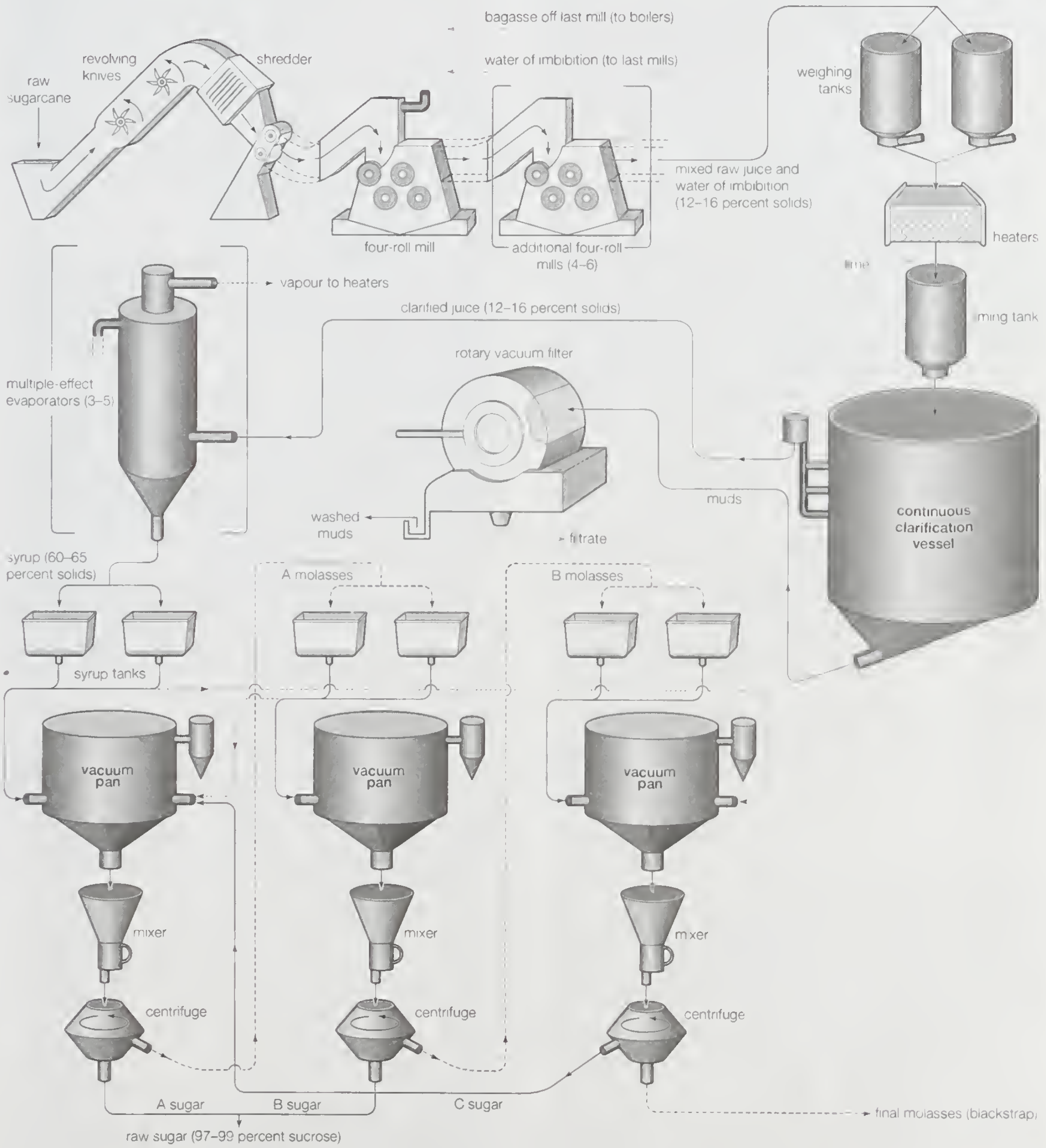

bagasse off last mill (to boilers)
revolving knives
shredder
raw sugarcane
water of imbibition (to last mills)
weighing tanks
mixed raw juice and water of imbibition (12–16 percent solids)
heaters
four-roll mill
additional four-roll mills (4–6)
lime
vapour to heaters
clarified juice (12–16 percent solids)
liming tank
rotary vacuum filter
multiple-effect evaporators (3–5)
muds
continuous clarification vessel
washed muds
filtrate
syrup (60–65 percent solids)
A molasses
B molasses
syrup tanks
vacuum pan
vacuum pan
vacuum pan
mixer
mixer
mixer
centrifuge
centrifuge
centrifuge
A sugar
B sugar
C sugar
final molasses (blackstrap)
raw sugar (97–99 percent sucrose)

Figure 32: Essential steps in the processing of sugarcane.
Encyclopædia Britannica, Inc.

often in a sucrose solution, which forms a calcium saccharate compound. The heat and lime kill enzymes in the juice and increase pH from a natural acid level of 5.0–6.5 to a neutral pH. Control of pH is important throughout sugar manufacture because sucrose inverts, or hydrolyzes, to its components glucose and fructose at acid pH (less than 7.0), and all three sugars decompose quickly at high pH (greater than 11.5).

Heated to 99°–104° C (210°–220° F), the neutralized juice is inoculated, if necessary, with flocculants such as polyacrylamides and pumped to a continuous clarification vessel, a large, enclosed, heated tank in which clear juice flows off the upper part while muds settle below. This settling and separation process is known as defecation. Muds are pumped to rotary vacuum filters, where residual sucrose is washed out with a water spray on a rotating filter. Clarified juice, meanwhile, is pumped to a series of three to five multiple-effect evaporators.

Multiple-effect evaporators

Concentration. In the multiple-effect system, developed for the American sugar industry in 1843, steam is used to heat the first of a series of evaporators. The juice is boiled and drawn to the next evaporator, which is heated by vapour from the first evaporator. The process continues through the series until the clarified juice, which consists of 10–15 percent sucrose, is concentrated to evaporator syrup, consisting of 55–59 percent sucrose and 60–65 percent by weight total solids. Nonsugars deposit on the walls and tubes of the evaporators, creating scale deposits and reducing the efficiency of heat transfer. Scale removal often forces the entire factory operation to shut down if another set of evaporators is not available.

Crystallization. Syrup from the evaporators is sent to vacuum pans, where it is further evaporated, under vacuum, to supersaturation. Fine seed crystals are added, and the sugar "mother liquor" yields a solid precipitate of about 50 percent by weight crystalline sugar. Crystallization is a serial process. The first crystallization, yielding A sugar or A strike, leaves a residual mother liquor known as A molasses. The A molasses is concentrated to yield a B strike, and the low-grade B molasses is concentrated to yield C sugar and final molasses, or blackstrap. Blackstrap contains approximately 25 percent sucrose and 20 percent invert (glucose and fructose); at these levels the sugar cannot be removed economically by crystallization.

Crystal separation and drying. Crystals and mother liquor are separated in basket-type centrifuges. Continuous machines are used for C and sometimes B sugars, but batch machines are best for A sugars because of the crystal breakage that takes place in continuous centrifuges. Mother liquor is spun off the crystals, and a fine jet of water is sprayed on the sugar pressed against the wall of the centrifugal basket, reducing the syrup coating on each crystal. In modern factories, the washing process is quite extensive in an effort to produce high-purity raw sugar. Overall recovery of sugar from cane juice averages between 70 and 80 percent.

The washed sugar, dumped from the baskets onto moving belts, dries and cools on the belts as it moves to bulk storage. At this point it is pale brown to golden yellow, with a sucrose content of 97–99 percent and a moisture content of 0.5 percent. This raw sugar, the sugar of commerce, is stored in bags in countries where labour is abundant and cheap. Generally, however, it is stored in bulk and shipped loose, like grain, in dry-bulk ships to areas where it will be refined.

Raw sugar products. *Open pan sugar.* In industrial sugarcane processing, crystallization is conducted under vacuum in order to lower operating temperatures, but some sugar is produced in the tropics by "open pan" processes. In these processes, crudely clarified juices are boiled down in opcn containers until a sludgy mass of crystals can be transferred to molds. The hardened brown product is sold as *panela* or *piloncillo* in Latin America and as *gur* or *khansari* in Asia.

Plantation white sugar. Plantation white, or mill white, sugar is a white sugar commonly produced for local consumption in sugarcane-growing countries. It is produced at the factory without remelting and refining of the raw sugar. Instead, sulfur dioxide gas (produced by burning sulfur in air) is injected into extracted juice, where it bleaches juice colorants, is oxidized to sulfate, and then is neutralized by the addition of lime. (Sulfite salts are sometimes substituted for sulfur dioxide.) A white sugar results that is suitable for table use but not for food processing, because it contains all the nonsugars (including bleached or reduced colorant) present in raw sugar.

Bleaching unrefined sugar

Higher grades of plantation white are produced by a carbonatation purification process, in which carbon dioxide gas (scrubbed flue gas) is injected into juice and reacted with lime to form calcium carbonate, which absorbs and adsorbs nonsugars and is filtered off. Powdered vegetable carbon is sometimes added to increase decolorization.

As demand for high-quality white sugars increases among food processors and beverage manufacturers in tropical areas, the processes described above are being improved and replaced by "Blanco Directo" processes, in which colour-precipitating reagents remove colorants instead of temporarily bleaching them.

Sugar refining. Sugar refining is the production of high-quality sugars from remelted raw cane sugars. ("Refining" is also used in beet sugar factories to describe the remelting and recrystallization processes by which high-quality white sugars are made from lower-grade beet syrups; see below *Beet sugar.*) About 35 percent of cane sugar is refined; the remainder is consumed as plantation white or as raw sugar. In tropical regions, small "white end" refineries are often built to refine the raw sugar produced by cane-processing plants. Raw sugar factories produce their own steam by burning bagasse, and a reasonably efficient plant has as much as 20 percent excess bagasse. This can be burned to operate the white end refinery, or it can be used to run a distillery or even generate electricity to be sold to the local power grid.

Still, most sugar refining is conducted in the consuming regions by large refineries, which produce a range of products such as soft brown sugars, sugar cubes, and granulated sugar. At these refineries, the raw sugar is affined (washed), melted (dissolved), sent through processes of clarification and decolorization, and crystallized. Sugar products are then dried, packaged, and stored.

Affination and melting. Affination is the mingling of raw sugar with a warm, heavy syrup, which removes the molasses coating from the sugar crystal. The syrup and crystals are separated in a spinning centrifugal basket, and the crystals are further "washed" by a water spray. Washed raw sugar is fed by screw conveyor to a melter, where it is dissolved at 65° C (150° F) in hot, sweet water with some fresh, hot water added to obtain a raw liquor containing about 65 percent dissolved solids.

Clarification and decolorization. Melt syrup is clarified either by phosphatation, in which phosphoric acid and lime are added to form calcium phosphates, which are removed by surface scraping in a flotation clarifier, or by carbonatation, in which carbon dioxide gas and lime form calcium carbonate, which is filtered off. Colour precipitants are added to each process.

The carbonatated liquors are filtered in pressure leaf filters with the use of diatomaceous earth, a filter aid invented for sugar processing. The resultant yellow to light brown liquor is further decolorized by carbon adsorbents, such as granular activated carbon or bone charcoal, or by ion-exchange resins of acrylic or styrenic materials. Decolorization is conducted in columns in various serial or parallel conformations.

Crystallization. Fine clarified liquor is boiled to white sugar in a series of vacuum pans similar to those used in sugarcane processing. The boiling system is complicated because the purity of the fine liquor is more than 98 percent, and at least six or seven stages of boiling are necessary before the molasses is exhausted. The first three or four strikes are blended to make commercial white sugar. Special large-grain sugar (for bakery and confectionery) is boiled separately. Fine grains (sanding or fruit sugars) are usually made by sieving products of mixed grain size. Powdered icing sugar, or confectioners' sugar, results when white granulated sugar is finely ground, sieved, and mixed with small quantities (3 percent) of starch or calcium phosphate to keep it dry. Brown sugars (light to dark) are

Large-grain, fine-grain, and powdered sugar either crystallized from a mixture of brown and yellow syrups (with caramel added for darkest colour) or made by coating white crystals with a brown-sugar syrup.

Packaging and storage. Crystalline and liquid sucrose products are dried and packaged in food-grade packaging plants. Package sizes range from individual servings to one-ton bags, packaging materials from paper to plastic-lined burlap or fabric. Sugar cubes are made by mixing white (or brown) sugar with syrup and then molding and drying.

Refineries can also produce syrups in a range of colours and flavours for the food-processing industry. Products include liquid sucrose as well as invert syrups (syrups containing all or partially inverted sucrose).

BEET SUGAR

Beet sugar factories generally produce only white sugar from sugar beets. Brown sugars are made with the use of cane molasses as a mother liquor component or as a crystal coating.

Sugar beet harvest and delivery. Sugar beets are grown in temperate areas of Europe, North America, and northern Asia. They are harvested from September through November, almost always by multirow harvester machines. The machines remove some dirt, the leaves, and sometimes the crown (depending on the contract terms). Because sugar does not deteriorate as severely in beets as it does in sugarcane shortly after harvest, a full crop of beets can be lifted (harvested) and stored for several weeks at ambient temperature or even for several months at freezing temperatures.

Beets are delivered by rail or road transport to the factory, where they are weighed in and sampled for analysis. Sampling schemes vary in complexity; beets are analyzed for trash, soil, sugar content, and (where beet quality is part of the contract) nitrogen and salt content. Sugar beet, being a root, has a much higher nitrogen content than sugarcane, and these nitrogen compounds can affect certain processing steps.

Payment is split along lines similar to lines for sugarcane payment, 60–65 percent going to the grower and 35–40 percent to the factory.

White sugar production. *Washing and extraction.* When harvested sugar beets are off-loaded at the factory, they are washed in a flume to remove rocks and dirt and then fed by gravity through a hopper to the slicing machine. There the roots are cut into "cossettes," V-shaped strips, three by four to seven centimetres in size (approximately one by two to three inches) in order to offer maximum surface area for extraction. Sugar extraction takes place in a multicell countercurrent diffuser. In order to minimize microbial growth and the use of biocide, temperatures are maintained above 75° C (167° F). Some 98 percent of the sugar is extracted to form what is known as diffusion juice, or raw juice.

Remaining beet pulp, discharged at over 90 percent moisture content, is pressed and dried. Pulp driers are a major energy consumer at the beet factory, which must purchase fuel since pulp cannot be burned and has a high market value as feed.

From raw juice to clear juice

Purification. Raw juice (containing 10 to 14 percent sucrose) is purified in a series of liming and carbonatation steps, often with filtration or thickening being conducted between the first and second carbonatation. One popular multistage system involves cold pre-liming followed by cold main liming, hot main liming, first carbonatation, filtration and mud recirculating, addition of heat and soda, second carbonatation, and filtration.

After carbonatation, sulfur dioxide is pumped through the juice in order to lower the pH level and reduce the colour. Beet processing is generally at pH levels slightly above 7. At low pH, invert sugar would form and react with nitrogen compounds to form colour, and, at high pH, alkaline destruction of sucrose and monosaccharides would occur.

Concentration and crystallization. After purification, the juice, now called clear or thin juice, is pumped to multiple-effect evaporators similar to those used in raw cane sugar manufacture. In the evaporators the juice is concentrated to thick juice (60–65 percent dissolved solids), which is mixed with remelted lower grades of sugar to form standard liquor. From this standard liquor, sugar is crystallized, usually in three stages. In all boiling systems, sugar obtained from the first stage is processed as a final product, while sugar from the second and third stages is remelted and recycled into another batch of thick juice.

Sugar is separated from mother liquor in basket centrifuges, and it is dried in either rotary louvred driers or fluidized-bed dryer-coolers.

Packaging and storing. Before packing, it is important that all sugar be cooled below 45° C (113° F). At higher temperatures it hardens in the bag or silo and can develop colour. Beet sugar factories store white sugar in silos during production and pack sugar year-round to meet the current market.

Molasses processing. In order to increase production at the beet sugar factory, molasses desugarization is practiced. One prominent desugarization process is ion exclusion, which separates compounds by their molecular weight and electrical charge. A fraction containing salts and high-molecular-weight colorants and saccharides comes first off the resin column; then comes a sucrose fraction, and then a betaine fraction (trimethylglycine, a component of sugar beets), which may be sold as a separate product. The sucrose fraction is recycled into thick juice to form standard liquor. Ion-exchange processes reduce residual sucrose in molasses to 9–14 percent and increase a factory's overall yield by 10 percent.

SUGAR BY-PRODUCTS

By-products of cane sugar and beet sugar production include fibre (from both cane and beet) and molasses (residual concentrated syrup from which no more sugar can economically be removed).

Fibre. *Sugar beet pulp.* Sugar beet pulp is used almost entirely for animal feed, mixed with molasses in loose or pellet form. Because of the higher nitrogen content of sugar beets, nitrogen (in the form of urea) need not be added, as it must when sugarcane bagasse is used for animal feed. Other uses for beet pulp are as edible fibre, for addition of soluble fibre to baked goods and processed foods, and for inclusion in paper manufacture.

Bagasse. Feed use for bagasse is relatively minor. The major use is as fuel for the cane factory, where one ton of dry bagasse is equivalent in energy value to two barrels of fuel oil. Freshly produced bagasse contains about 50 percent moisture and becomes drier on storage.

Bagasse is also widely used as filler for paper, fibreboard, and particleboard—especially in areas where wood is in short supply. Paper quality ranges from kraft-process brown paper through newsprint to glossy white.

Molasses. Molasses from both sugarcane and sugar beets is a major component of animal feed. Sugar beet molasses that has been subjected to desugarization contains reduced carbohydrate levels and may be blended with cane molasses.

Production of ethanol (ethyl alcohol) for industry and distilled spirits is common at most cane and beet factories. Rum is produced from cane molasses in the Western Hemisphere; beverage alcohol is produced from beet molasses in Europe. (M.A.C.)

Cocoa products

The cocoa bean is the seed of the cacao tree (*Theobroma cacao*), a tropical plant indigenous to the equatorial regions of the Americas. From the processed cocoa bean comes the fluid paste, or liquor, from which cocoa powder and chocolate are made. Chocolate is sold directly to the consumer as solid bars of eating chocolate, as packaged cocoa, and as baking chocolate. It is also used by confectioners as coating for candy bars and boxed or bulk chocolates, by bakery product manufacturers and bakers as coating for many types of cookies and cakes, and by ice-cream companies as coating for frozen novelties. Cocoa powders, chocolate liquor, and blends of the two are used in bulk to flavour various food products and to provide the flavours in such "chocolate" products as syrups, toppings, chocolate milk, prepared cake mixes, and pharmaceuticals.

HISTORY OF USE

The cacao tree was cultivated more than 3,000 years ago by the Maya, Toltec, and Aztec, who prepared a beverage from the bean (sometimes using it as a ceremonial drink) and also used the bean as a currency.

Early uses of cocoa

Columbus brought cocoa beans to Spain after his fourth voyage in 1502, and the Spanish conquistadores, arriving in Mexico in 1519, were introduced to a chocolate beverage by the Aztec. The Aztec beverage was made from sun-dried shelled beans, probably fermented in their pods. The broken kernels, or nibs, were roasted in earthen pots, then ground to paste in a concave stone, called a *metate*, over a small fire. Vanilla and various spices and herbs were added, and corn (maize) was sometimes used to produce milder flavour. The paste, formed into small cakes, was cooled and hardened on shiny leaves placed under a tree. The cakes were broken up, mixed with hot water, and beaten to foamy consistency with a small wooden beater, a molinet, producing the beverage called *xocoatl* (from Nahuatl words meaning "bitter water").

Too bitter for European taste, the mixture was sweetened with sugar when introduced to the Spanish court. Although Spain guarded the secret of its *xocoatl* beverage for almost 100 years, it reached Italy in 1606 and became popular in France with the marriage of the Spanish princess Maria Theresa to Louis XIV in 1660. In 1657 a Frenchman opened a London shop, selling solid chocolate to be made into the beverage, and chocolate houses, selling the hot beverage, soon appeared throughout Europe. By 1765 chocolate manufacture began in the American colonies at Dorchester, in Massachusetts, using cocoa beans from the West Indies.

In 1828 C.J. van Houten of The Netherlands patented a process for obtaining "chocolate powder" by pressing much of the cocoa butter from ground and roasted cocoa beans. In 1847 the English firm of Fry and Sons combined cocoa butter, a by-product of the pressing, with chocolate liquor and sugar to produce eating chocolate, and in 1876 Daniel Peter of Switzerland added dried milk to make milk chocolate. The proliferation of flavoured, solid, and coated chocolate foods rapidly followed.

Starting in the Americas in an area stretching from southern Mexico to the northern countries of South America, commercial cacao cultivation spread around the world to areas within 20° of the Equator where rainfall, temperatures, and soil conditions were suitable for its growth.

COCOA BEAN PROCESSING

Harvesting. Harvesting of cocoa beans can proceed all year, but the bulk of the crop is gathered in two flush periods occurring from October to February and from May to August. The ripe seed pods are cut from the trees and split open with machetes. The beans, removed from the pods with their surrounding pulp, are accumulated in leaf-covered heaps, in leaf-lined holes dug in the ground, or in large shallow boxes having perforated bottoms to provide for drainage.

Fermentation. The pulp of common grades (Forastero) is allowed to ferment for five to seven days, and the pulp of the more distinctively flavoured grades (Criollo) for one to three days. Frequent turnings dissipate excess heat and provide uniformity. During fermentation, the juicy sweatings of the pulp are drained away, the germ in the seed is killed by the increased heat, and flavour development begins. The beans become plump and full of moisture, and the interior develops a reddish brown colour and a heavy, sharp fragrance. The fermented beans are sun-dried or kiln-dried to reduce moisture content to 6–7 percent and bagged for shipment.

Cleaning, roasting, and grinding. Cocoa beans are subjected to various cleaning processes to remove such contaminants as twigs, stones, and dust. Roasting develops flavour, reduces acidity and astringency, lowers moisture content, deepens colour, and facilitates shell removal. After roasting comes a cracking and fanning (winnowing) process, in which machines crack the shells and then separate them from the heavier nibs by means of blowers. The cell walls of the nibs are in turn broken by grinding, releasing the fat, or cocoa butter, and forming a paste called chocolate liquor, or cocoa mass. If alkalized (Dutched) chocolate liquor is to be produced, the cocoa beans may be winnowed raw; the raw nibs will be alkalized and then roasted prior to grinding.

Conching. Conching, a flavour-developing, aerating, and emulsifying procedure performed by conche machines, requires from 4 to 72 hours, depending on the results desired and the machine type. Temperatures used in this process range from 55° to 88° C (130° to 190° F) and are closely controlled to obtain the desired flavour and uniformity.

Molding. In molding, the chocolate is cast in small consumer-size bars or in blocks weighing about 10 pounds for use by confectioners, then subjected to cold air to produce hardening.

COCOA BEAN PRODUCTS

Cocoa powders. Cocoa powders are produced by pulverizing cocoa cakes, made by subjecting the chocolate liquor of about 53 to 56 percent cocoa butter content to hydraulic pressing to remove a predetermined amount of cocoa butter. The cocoa butter content remaining in the powder may range from 8 to 36 percent, with the most common commercial grades in the United States containing 11, 17, or 22 percent cocoa butter. In the United Kingdom, cocoa sold for beverage use must contain a minimum of 20 percent.

Cocoa butter content

Natural process. Natural-process cocoa powders and chocolate liquors receive no alkali treatment. Cocoa beans are normally slightly acid, with a pH of 5.2–5.8. When the pH remains unchanged, the beans produce pleasantly sharp flavours blending well in many foods and confections.

Dutch process. Dutch-process cocoa powders and chocolate liquors are treated at the nib, liquor, or powder stage. The treatment is frequently referred to as "Dutching" because the process, first applied by C.J. van Houten in The Netherlands, was introduced as "Dutch cocoa." In this alkalizing process, a food-grade alkali solution may be applied in order partially to neutralize the natural cocoa acids, mostly acetic acid like that in vinegar; or it may be used to produce a strictly alkaline product, with a pH as high as 8.0. Potassium carbonate is most commonly used as an alkalizer, although other alkalies, such as sodium carbonate, may be used. In addition to altering the pH of the cocoa powder, the process darkens colour, mellows flavour, and alters taste characteristics.

Chocolate products. Chocolate products usually require the addition of more cocoa butter to that already existing in the chocolate liquor. The various forms of chocolate are available in consumer-size packages and in large bulk sizes for use by food manufacturers and confectioners. Most European confectioners make their own chocolate; other confectioners buy chocolate from chocolate-manufacturing specialists. For large commercial orders, chocolate is shipped, warm and in liquid form, in heated sanitary tank trucks or tank cars.

Baking chocolate. Baking (bitter) chocolate, popular for household baking, is pure chocolate liquor made from finely ground nibs, the broken pieces of roasted, shelled cocoa beans. This chocolate, bitter because it contains no sugar, can be either the natural or alkalized type.

Sweet chocolate. Sweet chocolate, usually dark in colour, is made with chocolate liquor, sugar, added cocoa butter, and such flavourings as vanilla beans, vanillin, salt, spices, and essential oils. Sweet chocolate usually contains at least 15 percent chocolate liquor content, and most sweet chocolate contains 25–35 percent. The ingredients are blended, refined (ground to a smooth mass), and conched. Viscosity is then adjusted by the addition of more cocoa butter, lecithin (an emulsifier), or a combination of both.

Milk chocolate. Milk chocolate is formulated by substituting whole milk solids for a portion of the chocolate liquor used in producing sweet chocolate. It usually contains at least 10 percent chocolate liquor and 12 percent whole milk solids. Manufacturers usually exceed these values, frequently going to 12–15 percent chocolate liquor and 15–20 percent whole milk solids. Milk chocolate, usually lighter in colour than sweet chocolate, is sweeter or milder in taste because of its lower content of bitter chocolate liquor. Processing is similar to that of sweet chocolate. "Bitter chocolate" refers to either baking chocolate or bittersweet chocolate. Bittersweet is similar to sweet chocolate but contains less sugar and more chocolate liquor. Minimum percentages of chocolate liquor are fixed by law in some countries, such as the United States.

Chocolate-type coatings. Confectionery coatings are made in the same manner as similar chocolate types, but some or all of the chocolate liquor is replaced with equivalent amounts of cocoa powder, and instead of added cocoa butter, with a melting point of about 32°–33° C (90°–92° F), other vegetable fats of equal or higher melting points are used. In the United States the legal name of this coating is "sweet cocoa and vegetable fat (other than cocoa fat) coatings." In the "chocolate" coating usually applied to ice cream and other frozen novelties, legally known as "sweet chocolate and vegetable fat (other than cocoa fat) coatings," the added cocoa butter usual in chocolate is replaced by lower-melting-point vegetable fats, such as coconut oil.

By-products. Shells, the major by-product of cocoa and chocolate manufacturing, represent 8–10 percent of raw cocoa bean weight and are blown off in the cracking and fanning, or winnowing, operation. They are used for fertilizer, mulch, and fuel.

Chocolate and cocoa grades. In chocolate and cocoa products, there is no sharp difference from one grade or quality to the next. Chocolate quality depends on such factors as the blend of beans used, with about 20 commercial grades from which to choose; the kind and amount of milk or other ingredients included; and the kind and degree of roasting, refining, conching, or other type of processing employed. Chocolate and cocoa products are only roughly classified; there are hundreds of variations on the market, alone or in combination with other foods or confections.

Care and storage. Chocolate and cocoa require storage at 18°–20° C (65°–68° F), with relative humidity below 50 percent. High (27°–32° C, or 80°–90° F) or widely fluctuating temperatures will causc fat bloom, a condition in which cocoa butter infiltrates to the surface, turning products gray or white as it recrystallizes.

Fat bloom and mustiness

High humidity causes mustiness in cocoa powder and can lead to mold formation in cocoa powder or on chocolate. Excessive moisture can also dissolve sugar out of chocolate, redepositing it on the surface as sugar bloom, distinguished from fat bloom by its sandy texture.

Nutritive value. Cocoa, a highly concentrated food providing approximately 1,000 calories per kilogram, provides carbohydrates, fat, protein, and minerals. Its theobromine and caffeine content produce a mildly stimulating effect. The carbohydrates and easily digested fats in chocolate make it an excellent high-energy food. (L.R.C./Ed.)

Confectionery products

The application of the terms confectionery and candy varies among English-speaking countries. In the United States candy refers to both chocolate products and sugar-based confections; elsewhere "chocolate confectionery" refers to chocolates, "sugar confectionery" to the various sugar-based products, and "flour confectionery" to such products as cakes and pastries. This section is primarily concerned with sugar confectionery. Other types of confections are discussed in the sections *Bakery products* and *Cocoa products.*

HISTORY

Egyptian hieroglyphics dating back at least 3,000 years indicate that the art of sugar confectionery was already established. The confectioner was regarded as a skilled craftsman by the Romans, and a confectioner's kitchen excavated at Herculaneum was equipped with pots, pans, and other implements similar to those in use today.

Early confectioners, not having sugar, used honey as a sweetener and mixed it with various fruits, nuts, herbs, and spices.

During the Middle Ages the Persians spread sugarcane cultivation, developed refining methods, and began to make a sugar-based candy. A small amount of sugar was available in Europe during the Middle Ages and was used in the manufacture of the confections prepared and sold mainly by apothecaries. The Venetians brought about a major change in candy manufacture during the 14th century, when they began to import sugar from Arabia. By the 16th century confectioners were manufacturing sweets by molding boiled sugar with fruits and nuts into fanciful forms by simple hand methods. The development of candy-manufacturing machinery began in the late 18th century.

INGREDIENTS

Sweeteners. Sugar, mainly sucrose from sugar beets or sugarcane, is the major constituent of most candies. Other sweeteners employed in candy manufacture include corn syrup, corn sugar, honey, molasses, maple sugar, and noncaloric sweeteners. Sweeteners may be used in dry or liquid form.

Invert sugar

Invert sugar, a mixture of glucose (dextrose) and fructose produced from sugar (sucrose) by application of heat and an acid "sugar doctor," such as cream of tartar or citric acid, affects the sweetness, solubility, and amount of crystallization in candymaking. Invert sugar is also prepared as a syrup of about 75 percent concentration by the action of acid or enzymes on sugar in solution.

Texturizers and flavourings. Because of the perishability of fresh fluid milk and milk products, milk is usually used in concentrated or dried form. It contributes to candy flavour, colour, and texture. Fats, usually of vegetable origin, are primarily used to supply textural and "mouth feel" propcrtics (lubrication and smoothness). They arc also used to control crystallization and to impart plasticity. Such colloids as gelatin, pectin, and egg albumin are employed as emulsifying agents, maintaining fat distribution and providing aeration. Other ingredients include fruits; nuts; natural, fortified, and artificial flavours; and colourings.

PRODUCTS

Candies can be divided into noncrystalline, or amorphous, and crystalline typcs. Noncrystallinc candies, such as hard candies, caramels, and toffces, are chcwy or hard, with homogeneous structure. Crystalline candies, such as fondant and fudgc, are smooth, creamy, and easily chewed, with a dcfinite structurc of small crystals.

High-boiled, or hard, candy. *Properties.* Sugar has the property of forming a type of noncrystallinc "glass" that forms the basis of hard candy products. Sugar and water are boiled until the concentration of the solution reaches a high level, and supersaturation persists upon cooling. This solution takes a plastic form and on further cooling becomes a hard, transparent, glassy mass containing lcss than 2 percent watcr.

Precisc control of boiling

High-boiled sugar solutions are unstable, however, and will readily crystallize unless prcventative stcps are taken. Control of modern sugar-boiling processes is precise. Crystallization is prcvcnted by adding either manufactured invert sugar or corn syrup. The latter is now favoured because it contains complex saccharides and dextrins that, in addition to increasing solubility, give greater viscosity, considerably retarding crystallization.

Hard candy manufacture. Originally, hard candy syrups were boiled ovcr a coke or gas fire. Modern manufacturers use pans jacketed with high-pressure steam for batch boiling. Special steam-pressure cookers through which syrup passes continuously are used when a constant supply is required. For flavouring and colouring, thc batch of boiled syrup is turned out on a table to cool. While still plastic, the ingredients are kncaded into the batch; this may be done mechanically. In continuous production, flavours may be added to the hot liquid syrup. Especially prepared "sealed" flavours are then required to prevent loss by evaporation.

After flavouring, the plastic mass is shaped by passing through rollers with impressions or through continuous forming machines that produce a "rope" of plastic sugar.

By feeding a soft filling into the rope as a core, "bonbons" are made.

A satinlike finish may be obtained by "pulling" the plastic sugar. This consists of stretching the plastic mass on rotating arms and at the same time repeatedly overlapping. With suitable ratios of sugar to corn syrup, pulling will bring about partial crystallization and a short, spongy texture will result.

Caramels and toffee. The manufacture of caramel and toffee resembles hard candy making except that milk and fat are added. Sweetened, condensed, or evaporated milk is usually employed. Fats may be either butter or vegetable oil, preferably emulsified with milk or with milk and some of the syrup before being added to the whole batch. Emulsifiers such as lecithin or glyceryl monostearate are particularly valuable in continuous processes. The final moisture content of toffee and particularly of caramels is higher than that of hard candy. Because milk and fat are present, the texture is plastic at normal temperatures. The action of heat on the milk solids, in conjunction with the sugar ingredients, imparts the typical flavour and colour to these candies. This process is termed caramelization.

Because caramel is plastic at lower temperatures than hard candy, it may be extruded. Machines eject the plastic caramel under pressure from a row of orifices; the resulting "ropes" are then cut into lengths. Under continuous manufacturing, all ingredients are metred in recipe quantities into a container that gives an initial boil. Then the mixed syrup is pumped first into a continuous cooker that reduces the moisture content to its final level, and finally into a temporary holding vessel in which increased caramelization occurs, permitting the flavour obtained by the batch process to be matched. The cooked caramel is then cooled, extruded, and cut.

Fondant. Fondant, the basis of most chocolate-covered and crystallized crèmes (which themselves are sometimes called "fondants"), is made by mechanically beating a solution supersaturated with sugar, so that minute sugar crystals are deposited throughout the remaining syrup phase. These form an opaque, white, smooth paste that can be melted, flavoured, and coloured. Syrup made from corn syrup and sugar is now generally used for fondant.

Continuous production of fondant

Fully mechanical plants produce a ton of fondant per hour. Syrup, produced in a continuous cooker, is delivered to a rotating drum (see Figure 33) that is cooled internally with water sprays. The cooled syrup is scraped from the drum and delivered to a beater consisting of a water-cooled, rectangular box fitted internally with rotating pegged spindles and baffles. This gives maximum agitation while the syrup is cooling, causing very fine sugar crystals to be deposited in the syrup phase. The crystals, together with a small amount of air entrapped by the beating, give the fondant its typical white opacity. The proportion of sugar to corn syrup in the base syrup usually ranges from 3:1 to 4:1. The moisture content of fondant ranges from 12 to 13 percent.

Mechanically prepared fondant can be reheated without complete solution of the sugar-crystal phase, and it will be sufficiently fluid to be cast into molds. At the same time colourings and flavourings—fruit pulp, jam, essential oils, etc.—may be added. Remelting is usually carried out in steam-jacketed kettles provided with stirrers at a temperature range between 65° and 75° C (145° and 155° F).

Shaped pieces of fondant for crystallizing or covering with chocolate are formed by pouring the hot, melted, flavoured fondant into impressions made in cornstarch. A shallow tray about two inches deep is filled with cornstarch, which is leveled off and slightly compressed. A printing board covered with rows of plaster, wood, or metal models of the desired shape is then pressed into the starch and withdrawn. Into these impressions the fondant is poured and left to cool. Next, the tray is inverted over a sieve; the starch passes through, leaving the fondant pieces on the sieve. After gentle brushing or blowing to remove adhering starch, the fondants are ready for covering or crystallizing. A machine known as a Mogul carries out all these operations automatically, filling trays with starch, printing them, depositing melted fondant, and stacking the filled trays into a pile. At the other end of the machine, piles of trays that contain cooled and set crèmes are unstacked and inverted over sieves, and the crèmes are removed to be brushed and air-blown. Empty trays are automatically refilled, and the cycle continues.

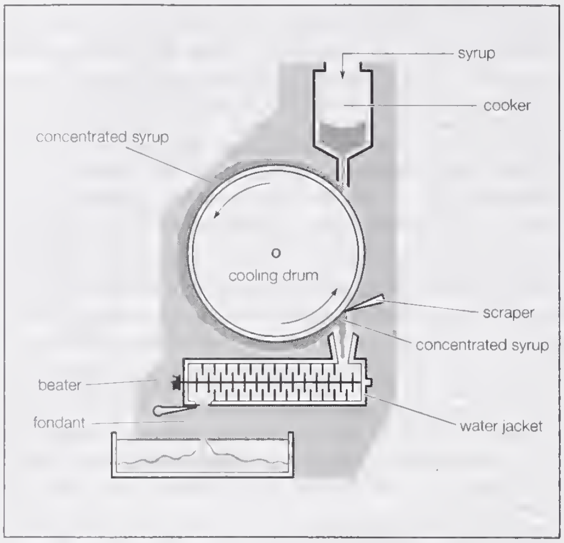


Figure 33: Continuous fondant-making machine.
Encyclopædia Britannica, Inc

Certain types of fondant may be remelted and poured into flexible rubber molds with impressions, but this process generally is limited to shallow crèmes of a fairly rigid consistency. Metal molds precoated with a substance that facilitates release of the crème also are used. The crème units are ejected from the inverted mold by compressed air onto a belt, which takes them forward for chocolate covering.

Fudge. Fudge combines certain properties of caramel with those of fondant. If hot caramel is vigorously mixed or if fondant is added to it, a smooth, crystalline paste forms on cooling. Known as fudge, this substance has a milky flavour similar to caramel and a soft, not plastic, texture. Fudge may be extruded or poured onto tables and cut into shapes. It is possible to construct a recipe that will pour into starch, but such fudge generally is inferior.

(H.B.K./B.W.M./Ed.)

Frozen prepared foods

Frozen prepared foods consist of complete meals or portions of meals that are precooked, assembled into a package, and frozen for retail sale. They are popular among consumers because they provide a diverse menu and are convenient to prepare. A typical frozen prepared meal contains a meat entree, a vegetable, a starch-based food such as pasta, and sauce. The manufacture of such a product requires careful attention by the food processor.

PREPARATION OF MEAL COMPONENTS

Meats. Meats are often one of the major components in a frozen meal. Several processing methods are used in preparing meats, such as marinating, cooking, and cutting or slicing.

Marinating. In order to tenderize the meat and develop desirable sensory attributes, a marinade is often helpful. Typical marinades contain salt, vinegar, lemon juice, spices, citric acid, and oil. Tenderization of meats is particularly enhanced by marinades that contain proteolytic enzymes—that is, enzymes that help to break down proteins. Meats are simply soaked in the marinade, or they are injected with marinade using special injection machines.

Cooking. Cooking of meats is necessary to eliminate all pathogens such as bacteria that produce harmful toxins. In a typical cooking process, the temperature at the centre of the meat is raised to 70° C (160° F) and held for at least two minutes.

Many meats are fried in immersion fryers. During frying, meats are cooked and desirable flavours created. Furthermore, the hot oil used in frying sears the surface of the meat, minimizing moisture loss during cooking. When meats are coated with breading material, frying is helpful in binding the batter. The oil retained in the breading layer enhances the aroma and texture of the fried foods.

Certain delicate foods, such as fish, are breaded and prefried for a short time to bind the breading material. Actual cooking of the fish is done when the consumer reheats the product. On the other hand, fried chicken is completely precooked during the frying process. Frozen fried chicken is reheated mainly to raise the serving temperature.

Commercial fryers

Commercial fryers are either batch or continuous units. In a continuous fryer, the foodstuff is placed on a feed conveyor that moves the product into a tank filled with frying oil. The oil is heated to 170°–180° C (340°–360° F). Simultaneously, another conveyor moves in the same direction just above the feed conveyor in order to prevent the food material from floating in the tank. The speed of the feed conveyor is carefully controlled so that the product remains immersed in oil for the required time. At the end of the tank, the conveyor moves the fried foodstuff above the oil, the surface oil being drained back into the tank.

Oven cooking is another method used to prepare main entrees in frozen prepared meals. Inside an oven, foods are heated by conduction, convection, or radiation. Certain ovens are designed to introduce steam during the heating cycle. In continuous-type ovens, the food moves on a mesh conveyor through different zones where the food may be subjected to different air velocities and steam flow in order to maintain the humidity at a desired level.

Batch-type ovens are ideally suited to cooking under vacuum. In vacuum cooking, meats are cooked at reduced pressure and temperature. In one vacuum technique, known as *sous-vide* cooking, foods are cooked in their own juices, thus retaining their natural flavours and moisture. Cooking time is usually increased because of the low temperatures employed. The process involves placing the food inside a laminated pouch and subjecting the packaged product to vacuum before sealing. The sealed pouch is then cooked in boiling water. After cooking, the pouches are quickly cooled in a stream of cold water prior to freezing.

Slicing and dicing. Cooked meats are more sensitive to physical handling than raw meats, because upon cooking the meat tissues become loosely connected with one another. Therefore, cooked meats are cooled to low temperatures, resulting in the stiffening of the muscle fibres, thus easing the cutting and slicing operations.

When the meats are in frozen state—that is, at temperatures between −18° and −23° C (0° and −10° F)—they are tempered before cutting. Tempering involves warming the frozen meats to temperatures slightly below their freezing point—for example, between −4° and −1° C (25° and 30° F). Tempering of frozen foods is often carried out in industrial-scale microwave ovens.

Meats are cut into cubes or dices by a dicing machine. A common industrial-scale dicer uses a knife blade attached to a revolving impeller. With each revolution of the impeller, the blade removes a slice from the large pieces of meat that are fed to the machine. The meat slices are cut into squares using cross-cut knives. The diced product is then discharged from the machine.

Vegetables. The vegetable portion of a prepared meal may be procured directly from a frozen-food processor, or raw vegetables may be frozen on site. The blanching and freezing of vegetables is described in *Vegetables: Processing of vegetables*. Processed vegetables that are intended for inclusion in prepared foods may be frozen if they are to be stored for a long duration, or they may be directly conveyed to the processing area for assembly on meal trays.

Pasta. In preparing pasta, size is an important criterion. Thin pastas such as spaghetti cook rapidly, so their texture is more difficult to control. Thick pastas, on the other hand, can be heated and cooled in a more controlled manner.

Pastas are cooked in rotary blanchers holding large volumes of water. A gentle rotary action in the blancher helps to avoid clumping during cooking. (Vegetable oil may also be added to the cooking water.) At the end of the cooking cycle, surface starch is washed off the pasta, again to avoid clumping. Immersing the cooked pasta in cold water also stops the cooking process.

Sauce. Sauces and gravies impart desirable sensory attributes to meats and vegetables. Furthermore, they help to prevent undesirable changes that result from dehydration of frozen foods during long-term storage. The most common sauces are either tomato- or cream-based. All sauces are accented with spices, thickening agents, emulsifiers, and salt.

Preparing ingredients. Dry ingredients are weighed and mixed in mixing blenders. The type of blender used depends on the physical characteristics of the ingredient particles and on whether any liquids or shortening agents are to be added to the mix. Complete mixing of the ingredients with the liquids or shortenings is vital to prevent inconsistencies in the final product.

Tomato and cream sauces

Tomato sauce is often made from tomato paste. Tomato paste usually contains from 24 to 36 percent tomato solids. Typically, it is procured in drums or flexible multiwall bags. Water is pumped in to flush out the paste and to help in diluting it to the desired concentration for sauce. The resulting tomato puree is then mixed with other ingredients to prepare the sauce.

Cream-based sauces begin with stock solutions, which are prepared by boiling raw stock material such as beef, fish, or poultry in water. Boiling is conducted in large kettles that may be operated either open to the atmosphere or under vacuum. Boiling under vacuum, accomplished at temperatures lower than 100° C (212° F), helps to retain more flavour compounds in the stock. Salt, spices, and herbs are added during the cooking process. After cooking is completed, the muscle tissue is removed.

Mixing, cooking, and cooling. In preparing sauces, dry mixes and liquids are blended to obtain a well-mixed slurry. The slurry is then fed into a cooking vessel, typically an open steam-jacketed kettle. Kettles may also be fitted with agitators, blenders, or scrapers. The agitator, located on the central axis, moves the product away from the heat-transfer surface of the kettle, an action that provides thorough cooking or cooling of the food material. Sometimes the scraper fingers of the agitator may move the product into the path of a secondary agitator for enhanced blending. Scraping and agitating help to minimize the building up of burned material on the heated surface.

During cooking of sauces, continued heating results in the swelling and gelation of the thickening agents and in the extraction of flavours from the seasoning agents. In order to retain more of the volatile and natural colour pigments, cooking may be done under vacuum.

Cream sauces are usually homogenized after heating. Homogenization ensures that fat globules will retain the small size necessary to stabilize the resulting emulsion. The scraping action must be done at high speed to prevent scorching the product.

After cooking, sauces are cooled rapidly to approximately 4° C (40° F). If the sauces contain only small or no particulates, then plate heat exchangers are used for cooling. In a plate heat exchanger, there is an indirect contact between the sauce and a cooling medium such as chilled water. A countercurrent flow arrangement between the sauce and the cooling medium assures high energy efficiency. After cooling, the sauce is stored in chilled holding tanks.

If the sauce formulation involves additional particulates—*e.g.,* diced mushrooms, raisins, or cooked meat—it is often desirable to precook the particulates and then blend them into sauce that has been already cooked and cooled. Although the particulates can be added prior to the cooking process, cooking them in sauce is usually avoided because of the adverse effects of heating on their texture.

ASSEMBLY AND FREEZING

Packages. Meal components are commonly assembled on trays made of aluminum foil, paperboard, plastic, or ceramic (see Figure 34). Special areas are stamped into the trays to create spaces for individual meal components. This type of package can be conveniently used directly by the consumer for reheating and then serving the food. In the case of boil-in-bag meals, the bag is made of laminated film. After a foodstuff is placed inside the bag, the air inside is evacuated, and the bag with its contents is frozen. The consumer simply places the bag with its contents in boiling water to prepare the meal.

Assembly. In order to assemble the prepared meal, adequate amounts of each meal component must be available. Often, meat portions, vegetables, and pasta are taken to the processing areas in large trays. Sauces are conveyed to the depositing areas by pumps. Depositing machines are then used automatically to weigh and place the required amounts of a given component directly onto the tray or into the bag.

© Bilderberg/The Stock Market

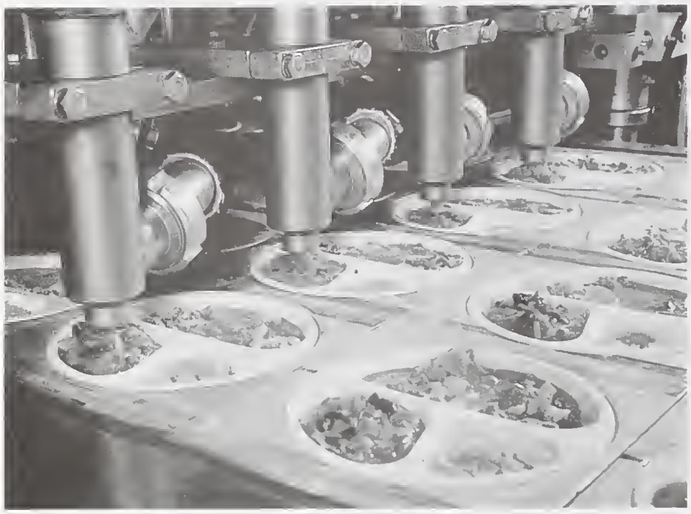


Figure 34: Depositing of prepared meals in trays prior to freezing.

Frozen vegetables are typically fed into a depositor through a feed chamber. The rotating bottom of the feed chamber usually contains several cavities that accept a designated amount of a food material. Under the depositing machines, the movement of the trays on the conveyor belts is sequenced. The cavities open directly over the meal tray above a designated slot. After the meat or vegetable is deposited on the tray, the sauce may be dispensed to coat the desired meal component.

Trays with meal components are inspected to assure that each item has been properly placed in the tray. Machines are used to detect any undesirable items such as metal or glass fragments. The trays are then conveyed directly to a freezer.

Freezing. Freezing of prepared and packaged meals is done rapidly to minimize changes in quality. Typically, once inside a freezer, foods are frozen to at least −40° C (−40° F) within 90 minutes.

Belt freezers

In one common type of freezer, the belt freezer, food trays or boil-in-bags are placed on a simple wire-mesh belt. The belt conveys the product into an air-blast room operated at −40° C. While a single belt arrangement is simple, a multitiered belt may be used to save floor space. In this case, a feed conveyor moves the product through several tiers of belts located inside the air-blast room.

Table 12: Processing Additives and Their Uses

function	typical chemical agent	typical product
Anticaking	sodium aluminosilicate	salt
Bleaching	benzoyl peroxide	flour
Chelating	ethylenediaminetetraacetic acid (EDTA)	dressings, mayonnaise, sauces, dried bananas
Clarifying	bentonite, proteins	fruit juices, wines
Conditioning	potassium bromate	flour
Emulsifying	lccithin	ice cream, mayonnaise, bakery products
Leavening	yeast, baking powder, baking soda	bakery products
Moisture control (humectants)	glyccrol	marshmallows, soft candies, chewing gum
pH control	citric acid, lactic acid	certain cheeses, confections, jams and jellies
Stabilizing and thickening	pectin, gelatin, carrageenan, gums (arabic, guar, locust bean)	dressings, frozen desserts, confections, pudding mixes, jams and jellies

A more compact arrangement employs a spiral belt. The spiral arrangement maximizes the belt surface area for a given floor space. A popular type of spiral freezer uses self-stacking belts. In the self-stacking arrangement, each tier rests on the vertical side flanges of the tier beneath. Several configurations of air flow are possible. Countercurrent vertical air flow, for instance, permits greater energy efficiencies. The air is channeled between the belts to minimize the time required for freezing. Faster rates of freezing minimize the dehydration of foods.

Plate freezers are commonly used for freezing brick-shaped packaged products. In plate freezers, refrigerant is allowed to circulate inside thin channels within the plates. The packaged products are firmly pressed between the plates. High rates of heat transfer can be obtained between the packaged product and the refrigerant plates.

After the freezing process, the frozen prepared foods are packaged in cartons. The cartons are labeled and stored in a frozen warehouse until needed for shipment to retail outlets. (R.P.Si.)

Food additives

Food additives, such as salt, spices, and sulfites, have been used since ancient times to preserve foods and make them more palatable. With the increased processing of foods in the 20th century, there came a need for both the greater use of and new types of food additives. Many modern products, such as low-calorie, snack, and ready-to-eat convenience foods, would not be possible without food additives.

Food additives and their metabolites are subjected to rigorous toxicological analysis prior to their approval for use in the industry. Feeding studies are carried out using animal species (*e.g.*, rats, mice, dogs) in order to determine the possible acute, short-term and long-term toxic effects of these chemicals. These studies monitor the effects of the compounds on the behaviour, growth, mortality, blood chemistry, organs, reproduction, offspring, and tumour development in the test animals over a 90-day to two-year period. The lowest level of additive producing no toxicological effects is termed the no-effect level (NOEL). The NOEL is generally divided by 100 to determine a maximum acceptable daily intake (ADI).

There are four general categories of food additives: nutritional additives, processing agents, preservatives, and sensory agents. These are not strict classifications, as many additives fall into more than one category.

NUTRITIONAL ADDITIVES

Nutritional additives are utilized for the purpose of restoring nutrients lost or degraded during production, fortifying or enriching certain foods in order to correct dietary deficiencies, or adding nutrients to food substitutes. The fortification of foods began in 1924 when iodine was added to table salt for the prevention of goitre. Vitamins are commonly added to many foods in order to enrich their nutritional value. For example, vitamins A and D are added to dairy and cereal products, several of the B vitamins are added to flour, cereals, baked goods, and pasta, and vitamin C is added to fruit beverages, cereals, dairy products, and confectioneries. Other nutritional additives include the essential fatty acid linoleic acid, minerals such as calcium and iron, and dietary fibre.

PROCESSING AGENTS

A number of agents are added to foods in order to aid in processing or to maintain the desired consistency of the product. Table 12 shows the functions performed by various processing agents employed in the food industry. Several of these agents are discussed in more detail below. For a discussion of leavening agents, see above *Bakery products: Ingredients: Leavening agents.*

Emulsifiers

Emulsifiers are used to maintain a uniform dispersion of one liquid in another, such as oil in water. The basic structure of an emulsifying agent includes a hydrophobic portion, usually a long-chain fatty acid, and a hydrophilic portion that may be either charged or uncharged. The hydrophobic portion of the emulsifier dissolves in the oil phase and the hydrophilic portion dissolves in the aqueous phase, forming a dispersion of small oil droplets. Emulsifiers thus form and stabilize oil-in-water emulsions (*e.g.,* mayonnaise), uniformly disperse oil-soluble flavour compounds throughout a product, prevent large ice crystal formation in frozen products (*e.g.,* ice cream), and improve the volume, uniformity, and fineness of baked products.

Stabilizers and thickeners have many functions in foods. Most stabilizing and thickening agents are polysaccharides, such as starches or gums, or proteins, such as gelatin. The primary function of these compounds is to act as thickening or gelling agents that increase the viscosity of the final product. These agents stabilize emulsions, either by adsorbing to the outer surface of oil droplets or by increasing the viscosity of the water phase. Thus, they prevent the coalescence of the oil droplets, promoting the separation of the oil phase from the aqueous phase (*i.e.,* creaming). The formation and stabilization of foam in a food product occurs by a similar mechanism, except that the oil phase is replaced by a gas phase. The compounds also act to inhibit the formation of ice or sugar crystals in foods and can be used to encapsulate flavour compounds.

Chelating, or sequestering, agents protect food products from many enzymatic reactions that promote deterioration during processing and storage. These agents bind to many of the minerals that are present in food (*e.g.,* calcium and magnesium) and are required as cofactors for the activity of certain enzymes.

PRESERVATIVES

Food preservatives are classified into two main groups: antioxidants and antimicrobials, as shown in Table 13. Antioxidants are compounds that delay or prevent the deterioration of foods by oxidative mechanisms. Antimicrobial agents inhibit the growth of spoilage and pathogenic microorganisms in food.

Antioxidants. The oxidation of food products involves the addition of an oxygen atom to or the removal of a hydrogen atom from the different chemical molecules found in food. Two principal types of oxidation that contribute to food deterioration are autoxidation of unsaturated fatty acids (*i.e.,* those containing one or more double bonds between the carbon atoms of the hydrocarbon chain) and enzyme-catalyzed oxidation.

The autoxidation of unsaturated fatty acids involves a reaction between the carbon-carbon double bonds and molecular oxygen (O_2). The products of autoxidation, called free radicals, are highly reactive, producing compounds that cause the off-flavours and off-odours characteristic of oxidative rancidity. Antioxidants that react with the free radicals (called free radical scavengers) can slow the rate of autoxidation. These antioxidants include the naturally occurring tocopherols (vitamin E derivatives) and the synthetic compounds butylated hydroxyanisole (BHA), butylated hydroxytoluene (BHT), and tertiary butylhydroquinone (TBHQ).

Prevention of browning

Specific enzymes may also carry out the oxidation of many food molecules. The products of these oxidation reactions may lead to quality changes in the food. For example, enzymes called phenolases catalyze the oxidation of certain molecules (*e.g.,* the amino acid tyrosine) when fruits and vegetables, such as apples, bananas, and potatoes, are cut or bruised. The product of these oxidation reactions, collectively known as enzymatic browning, is a dark pigment called melanin. Antioxidants that inhibit enzyme-catalyzed oxidation include agents that bind free oxygen (*i.e.,* reducing agents), such as ascorbic acid (vitamin C), and agents that inactivate the enzymes, such as citric acid and sulfites.

Antimicrobials. Antimicrobials are most often used with other preservation techniques, such as refrigeration, in order to inhibit the growth of spoilage and pathogenic microorganisms. Sodium chloride (NaCl), or common salt, is probably the oldest known antimicrobial agent. Organic acids, including acetic, benzoic, propionic, and sorbic acids, are used against microorganisms in products with a low pH. Nitratcs and nitrites are used to inhibit the bacterium *Clostridium botulinum* in cured meat products (*e.g.,* ham and bacon). Sulfur dioxide and sulfites are used to control the growth of spoilage microorganisms in dried fruits, fruit juices, and wines. Nisin and natamycin are preservatives produced by microorganisms. Nisin inhibits the growth of some bacteria while natamycin is active against molds and yeasts.

SENSORY AGENTS

Colorants. Colour is an extremely important sensory characteristic of foods; it directly influences the perception of both the flavour and quality of a product. The processing of food can cause degradation or loss of natural pigments in the raw materials. In addition, some formulated products, such as soft drinks, confections, ice cream, and snack foods, require the addition of colouring agents. Colorants are often necessary to produce a uniform product from raw materials that vary in colour intensity. Colorants used as food additives are classified as natural or synthetic. Natural colorants are derived from plant, animal, and mineral sources, while synthetic colorants are primarily petroleum-based chemical compounds.

Natural colorants. Most natural colorants are extracts derived from plant tissues. The use of these extracts in the food industry has certain problems associated with it, including the lack of consistent colour intensities, instability upon exposure to light and heat, variability of supply, reactivity with other food components, and addition of secondary flavours and odours. In addition, many are insoluble in water and therefore must be added with an emulsifier in order to achieve an cven distribution throughout the food product. Table 14 lists several natural colorants derived from plant extracts and used in various food products.

Table 13: Food Preservatives

chemical agent	mechanism of action
Antioxidants	
Ascorbic acid	oxygen scavenger
Butylated hydroxyanisole (BHA)	free radical scavenger
Butylated hydroxytoluene (BHT)	free radical scavenger
Citric acid	enzyme inhibitor/metal chelator
Sulfites	enzyme inhibitor/oxygen scavenger
Tertiary butylhydroquinone (TBHQ)	free radical scavenger
Tocopherols	free radical scavenger
Antimicrobials	
Acetic acid	disrupts cell membrane function (bacteria, yeasts, some molds)
Benzoic acid	disrupts cell membrane function/inhibits enzymes (molds, yeasts, some bacteria)
Natamycin	binds sterol groups in fungal cell membrane (molds, yeasts)
Nisin	disrupts cell membrane function (gram-positive bacteria, lactic acid-producing bacteria)
Nitrates, nitrites	inhibits enzymes/disrupts cell membrane function (bacteria, primarily *Clostridium botulinum*)
Propionic acid	disrupts cell membrane function (molds, some bacteria)
Sorbic acid	disrupts cell membrane function/inhibits enzymes/inhibits bacterial spore germination (yeasts, molds, some bacteria)
Sulfites and sulfur dioxide	inhibits enzymes/forms addition compounds (bacteria, yeasts, molds)

Numbering systems for colorants

Synthetic colorants. Synthetic colorants are water-soluble and are available commercially as powders, pastes, granules, or solutions. Special preparations called lakes are formulated by treating the colorants with aluminum hydroxide. They contain approximately 10 to 40 percent of the synthetic dye and are insoluble in water and organic solvents. Lakes are ideal for use in dry and oil-based products. The stability of synthetic colorants is affected by light, heat, pH, and reducing agents. A number of dyes have been chemically synthesized and approved for usage in various countries. These colorants are designated according to special numbering systems specific to individual countries. For example, the United States uses FD&C numbers (chemicals approved for use in foods, drugs, and cosmetics), and the European Union (EU) uses

Table 14: Natural Food Colorants

chemical class	colour	plant source	pigment	products
Anthocyanins	red	strawberry (*Fragaria* species)	pelargonidin 3-glucoside*	beverages, confections, preserves, fruit products
	blue	grape (*Vitis* species)	malvidin 3-glucoside*	beverages
Betacyanins	red	beetroot (*Beta vulgaris*)	betanin	dairy products, desserts, icings
Carotenoids**	yellow/orange	annatto (*Bixa orellana*)	bixin	dairy products, margarine
	yellow	saffron (*Crocus sativus*)	crocin	rice dishes, bakery products
	red/orange	paprika (*Capsicum annuum*)	capsanthin	soups, sauces
	orange	carrot (*Daucus carota*)	β-carotene	bakery products, confections
	red	mushroom (*Cantharellus cinnabarinus*)	canthaxanthin	sauces, soups, dressings
Phenolics	orange/yellow	turmeric (*Cuycuma longa*)	curcumin	dairy products, confections

*Plus other similar compounds. **Many carotenoids used as food colorants are chemically synthesized.

E numbers. Table 15 shows the most commonly used synthetic dyes.

All synthetic colorants have undergone extensive toxicological analysis. Brilliant Blue FCF, Indigo Carmine, Fast Green FCF, and Erythrosine are poorly absorbed and show little toxicity. Extremely high concentrations (greater than 10 percent) of Allura Red AC cause psychotoxicity, and Tartrazine induces hypersensitive reactions in some persons. Although none of the synthetic colorants listed above has been found to be carcinogenic in laboratory animals when administered orally, they are not universally approved in all countries. For example, while Allura Red AC is used extensively in the United States, it is banned from use in Canada.

Table 15: Synthetic Food Colorants

common name	designation		products
	United States	European Union	
Allura Red AC	FD&C Red No. 40	. . .	gelatin, puddings, dairy products, confections, beverages
Brilliant Blue FCF	FD&C Blue No. 1	E133	beverages, confections, icings, syrups, dairy products
Erythrosine	FD&C Red No. 3	E127	maraschino cherries
Fast Green FCF	FD&C Green No. 3	. . .	beverages, puddings, ice cream, sherbet, confections
Indigo Carmine	FD&C Blue No. 2	E132	confections, ice cream, bakery products
Sunset Yellow FCF	FD&C Yellow No. 6	E110	bakery products, ice cream, sauces, cereals, beverages
Tartrazine	FD&C Yellow No. 5	E102	beverages, cereals, bakery products, ice cream, sauces

Flavourings. The flavour of food results from the stimulation of the chemical senses of taste and smell by specific food molecules. Taste reception is carried out in specialized cells located in the taste buds. The four basic taste sensations—sweet, salty, bitter, and sour—are detected in separate regions of the tongue, mouth, and throat because the taste cells in each region are specific for certain flavour molecules (*e.g.*, sweeteners; see below).

In addition to the four basic tastes, the flavouring molecules in food stimulate specific olfactory (smell) cells in the nasal cavity. These cells can detect more than 10,000 different stimuli, thus fine-tuning the flavour sensation of a food.

A flavour additive is a single chemical or blend of chemicals of natural or synthetic origin that provides all or part of the flavour impact of a particular food. These chemicals are added in order to replace flavour lost in processing and to develop new products. Flavourings are the largest group of food additives, with more than 1,200 compounds available for commercial use. Natural flavourings are derived or extracted from plants, spices, herbs, animals, or microbial fermentations. Artificial flavourings are mixtures of synthetic compounds that may be chemically identical to natural flavourings. Artificial flavourings are often used in food products because of the high cost, lack of availability, or insufficient potency of natural flavourings.

Flavour enhancement

Flavour enhancers are compounds that are added to a food in order to supplement or enhance its own natural flavour. The concept of flavour enhancement originated in Asia, where cooks added seaweed to soup stocks in order to provide a richer flavour to certain foods. The flavour-enhancing component of seaweed was identified as the amino acid L-glutamate, and monosodium glutamate (MSG) became the first flavour enhancer to be used commercially. The rich flavour associated with L-glutamate was called *umami*. *Umami* is often considered the fifth basic taste because it is distinctly different from the other basic tastes (sweet, salty, sour, and bitter) and it is believed to activate a separate set of taste receptors.

Other compounds that are used as flavour enhancers include the 5′-ribonucleotides, inosine monophosphate (IMP), guanosine monophosphate (GMP), yeast extract, and hydrolyzed vegetable protein. Flavour enhancers may be used in soups, broths, sauces, gravies, flavouring and spice blends, canned and frozen vegetables, and meats.

Sweeteners. Sucrose or table sugar is the standard on which the relative sweetness of all other sweeteners is based. (For a discussion of sugar manufacture and refining, see above *Sugar*.) Because sucrose provides energy in the form of carbohydrates, it is considered a nutritive sweetener. Other nutritive sweeteners include glucose, fructose, corn syrup, high fructose corn syrup, and sugar alcohols (*e.g.*, sorbitol, mannitol, and xylitol).

Efforts to chemically synthesize sweeteners began in the late 1800s with the discovery of saccharin. Since then, a number of synthetic compounds have been developed that provide few or no calories or nutrients in the diet and are termed nonnutritive sweeteners. These sweeteners have significantly greater sweetening power than sucrose, and therefore a relatively low concentration may be used in food products. In addition to saccharin, the most commonly used nonnutritive sweeteners are cyclamates, aspartame, and acesulfame K. A comparison of the properties of these sweeteners is shown in Table 16.

The sensation of sweetness is transmitted through specific protein molecules, called receptors, located on the surface of specialized taste cells. All sweeteners function by binding to these receptors on the outside of the cells. The increased sweetness of the nonnutritive sweeteners relative to sucrose may be due to either tighter or longer binding of these synthetic compounds to the receptors.

Nonnutritive sweeteners are primarily used for the production of low-calorie products including baked goods, confectioneries, dairy products, desserts, preserves, soft

Table 16: Nonnutritive Sweeteners

chemical compound	relative sweetness (sucrose = 1)	temperature and pH stability	aftertaste
Acesulfame K	130–200	yes	bitter
Aspartame	150–200	no*	none
Cyclamates	30–80	yes	bitter, salty
Saccharin	200–700	yes	bitter, metallic

*An encapsulated preparation of aspartame has been developed for use in bakery products.

drinks, and tabletop sweeteners. They are also used as a carbohydrate replacement for persons with diabetes and in chewing gum and candies to prevent dental caries (*i.e.*, tooth decay). Unlike nutritive sweeteners, nonnutritive sweeteners do not provide viscosity or texture to products, so that bulking agents such as polydextrose are often required for manufacture.

Evaluating toxic effects of sweeteners

Toxicological analysis of the nonnutritive sweeteners has produced variable results. High concentrations of saccharin and cyclamates in the diets of rats have been shown to induce the development of bladder tumours in the animals. Because of these results, the use of cyclamates has been banned in several countries, including the United States, and the use of saccharin must include a qualifying statement regarding its potential health risks. However, no evidence of human bladder cancer has been reported with the consumption of these sweeteners. Both aspartame and acesulfame K are relatively safe, with no evidence of carcinogenic potential in animal studies. (P.M.D.)

BIBLIOGRAPHY

General works. R. MACRAE, R.K. ROBINSON, and M.J. SADLER (eds.), *Encyclopaedia of Food Science, Food Technology, and Nutrition,* 8 vol. (1993); and Y.H. HUI (ed.), *Encyclopedia of Food Science and Technology,* 4 vol. (1992), cover all aspects of the science of food. P. FELLOWS, *Food Processing Technology: Principles and Practices* (1988), is an introductory text. NORMAN N. POTTER, *Food Science,* 5th ed. (1995), treats the processing operations used for numerous food commodities. R. PAUL SINGH and FERNANDA A.R. OLIVEIRA (eds.), *Minimal Processing of Foods and Process Optimization* (1994), describes emerging technologies in this area.

R. PAUL SINGH and DENNIS R. HELDMAN, *Introduction to Food Engineering,* 2nd ed. (1993), provides introductory mathematical procedures useful in designing and analyzing food-processing operations. ROMEO T. TOLEDO, *Fundamentals of Food Process Engineering,* 2nd ed. (1991), an introductory text, contains mathematical analyses of a variety of food-processing operations. DENNIS R. HELDMAN and R. PAUL SINGH, *Food Process Engineering,* 2nd ed. (1981), an advanced text, quantitatively analyzes several unit operations commonly used in the food industry.

Food preservation and storage. MARCUS KAREL, OWEN R. FENNEMA, and DARYL B. LUND, *Physical Principles of Food Preservation* (1975), contains a quantitative description of commonly used food-processing operations. S.D. HOLDSWORTH, *Aseptic Processing and Packaging of Food Products* (1992), is a comprehensive text. JAMES M. JAY, *Modern Food Microbiology,* 4th ed. (1992), offers an introduction to the role of microorganisms in the food supply, covering the history of food microbiology as a science, the factors that affect microbial growth, the incidence and types of microorganisms found in foods, food preservation, and the part microorganisms play in food spoilage and related diseases. P.R. HAYES, *Food Microbiology and Hygiene,* 2nd ed. (1992), details the fundamentals of food microbiology and the hygienic aspects of the design and operation of food-processing equipment.

C.M.D. MAN and A.A. JONES (eds.), *Shelf Life Evaluation of Foods* (1994), examines various food commodities in terms of their shelf life and discusses methods used to study the shelf life of foods. THEODORE P. LABUZA, *The Shelf-Life Dating of Foods* (1982), a reference book, contains information on the shelf life of numerous food products and provides approaches to mathematical prediction of the shelf life of foods. (R.P.Si.)

Cereals and other starch products. The composition, processing, and uses of cereals and other starch products are discussed in GIUSEPPE FABRIANI and CLAUDIA LINTAS (eds.), *Durum Wheat: Chemistry and Technology* (1988); Y. POMERANZ (ed.), *Wheat: Chemistry and Technology,* 3rd ed., 2 vol. (1988), and *Wheat Is Unique: Structure, Composition, Processing, End-Use Properties, and Products* (1989); Y. POMERANZ, *Industrial Uses of Cereals* (1973), and *Modern Cereal Science and Technology* (1987); *Advances in Cereal Science and Technology* (biennial); STANLEY A. WATSON and PAUL E. RAMSTAD (eds.), *Corn: Chemistry and Technology* (1987); BIENVENIDO O. JULIANO (ed.), *Rice: Chemistry and Technology,* 2nd ed. (1985); RUTH H. MATTHEWS (ed.), *Legumes: Chemistry, Technology, and Human Nutrition* (1989); ALLAN K. SMITH and SIDNEY J. CIRCLE (eds.), *Soybeans: Chemistry and Technology,* vol. 1, *Proteins,* rev. ed. (1978); HARRY E. SNYDER and T.W. KWON, *Soybean Utilization* (1987); and SOY PROTEIN COUNCIL, *Soy Protein Products: Characteristics, Nutritional Aspects, and Utilization* (1987).

Some starch products are treated in R. GORDON BOOTH (ed.), *Snack Food* (1990), discussing the snack food industry in the United States and Great Britain; and ROBERT B. FAST and ELWOOD F. CALDWELL (eds.), *Breakfast Cereals, and How They Are Made* (1990). (Ed.)

Edible fats and oils. D.K. SALUNKHE *et al.*, *World Oilseeds: Chemistry, Technology, and Utilization* (1992), provides an overview. Nuts are treated in JASPER GUY WOODROOF, *Tree Nuts: Production, Processing, Products,* 2nd ed. (1979); and FREDERIC ROSENGARTEN, JR., *The Book of Edible Nuts* (1984). (Ed.)

Bakery products. Various aspects of the topic are treated in books by SAMUEL A. MATZ: *Ingredients for Bakers* (1987), containing descriptions of the raw materials used by bakers, *Formulas and Processes for Bakers* (1987), presenting guidelines used in formulating doughs and batters, examples of many typical formulas, explanations of changes occurring in mixing, fermenting, shaping, and cooking, and the effects that processing variables may have on the quality of the finished product, *Equipment for Bakers* (1988), discussing the different machines used in retail and wholesale bakeries, including the specifications and functions of individual models, *Bakery Technology: Packaging, Nutrition, Product Development, QA* (1989), and *Cookie and Cracker Technology,* 3rd ed. (1992). Additional works include E.J. PYLER, *Baking Science & Technology,* 3rd ed., 2 vol. (1988); WILLIAM J. SULTAN, *Practical Baking,* 5th ed. (1990), a survey of the art and craft of professional baking; and D.J.R. MANLEY, *Technology of Biscuits, Crackers, and Cookies,* 2nd ed. (1991).

A more fundamental approach to some of the problems in baking science can be found in HAMED A. FARIDI and JON M. FAUBION (eds.), *Dough Rheology and Baked Product Texture* (1990). (S.A.M.)

Fruits. A.C. HULME (ed.), *The Biochemistry of Fruits and Their Products,* 2 vol. (1970–71), provides a detailed discussion, as does the work by Jen cited in the paragraph below. DONALD L. DOWNING, *Processed Apple Products* (1989), the definitive text on this particular topic, details all significant commercial processes, including equipment and procedures. R.M. SMOCK and A.M. NEUBERT, *Apples and Apple Products* (1950), offers an extraordinary collection of bibliographic references. (M.R.McL.)

Vegetables. JOSEPH J. JEN (ed.), *Quality Factors of Fruits and Vegetables: Chemistry and Technology* (1989), describes the four major factors—colour, flavour, texture, and nutritive value—that determine food quality and discusses new technology used in food processing. Other works on vegetable processing include BOR SHIUN LUH and JASPER GUY WOODROOF, *Commercial Vegetable Processing,* 2nd ed. (1988), the most complete treatment of the subject, with chapters on quality control, nutrition labeling, and computer usage in food processing; WILLIAM F. TALBURT and ORA SMITH (eds.), *Potato Processing,* 4th ed. (1987), a comprehensive text; and MAS YAMAGUCHI, *World Vegetables: Principles, Production, and Nutritive Values* (1983), presenting a global view of vegetable production. (Jo.J.J.)

Fish. GEORGE M. PIGOTT and BARBEE W. TUCKER, *Seafood: The Effects of Technology on Nutrition* (1990), aimed at a general audience, discusses the effects that handling and processing methods may have on the nutritional value of seafood. E. GRAHAM BLIGH (ed.), *Seafood Science and Technology* (1992), written for the professional, covers the basic principles of seafood chemistry, microbiology, and technology. Two articles by GEORGE M. PIGOTT, "Flavors and Acceptance of Formulated Seafood Products," *Food Reviews International,* 6(4):661–680 (1990), and "Who Is the 21st Century Consumer?" *INFOFISH International,* 1:12–20 (January–February 1994), are also useful. (G.M.P.)

Meat. Detailed overviews include HAROLD B. HEDRICK *et al.*, *Principles of Meat Science,* 3rd ed. (1994); JOHN R. ROMANS *et al.*, *The Meat We Eat,* 13th ed. (1994); and H.R. CROSS and A.J. OVERBY (eds.), *Meat Science, Milk Science, and Technology* (1988), which includes comparisons of methods of meat production and processing in various countries.

PETER J. BECHTEL (ed.), *Muscle As Food* (1986); and A.J. BAILEY and N.D. LIGHT, *Connective Tissue in Meat and Meat Products* (1989), discuss physical and biochemical aspects.

Processed-meat science and technology is treated in HERBERT W. OCKERMAN, *Sausage and Processed Meat Formulations* (1989); and A.M. PEARSON and F.W. TAUBER, *Processed Meats*, 2nd ed. (1984). HERBERT W. OCKERMAN and C.L. HANSEN, *Animal By-Product Processing* (1988), covers the production of edible meat products, hides, glue, bone and meat meals, pharmaceutical products, sausage casings, pet foods, and animal waste products. (H.R.C.)

Poultry. ROBERT E. MORENG and JOHN S. AVENS, *Poultry Science and Production* (1985, reissued 1991), is a well-illustrated overview of all aspects of the poultry industry and avian biology. WILLIAM J. STADELMAN *et al., Egg and Poultry-Meat Processing* (1988), deals with nutritional aspects and contains a full listing of USDA-approved poultry products. (J.M.Re.)

Eggs. WILLIAM J. STADELMAN and OWEN J. COTTERILL, *Egg Science and Technology*, 3rd ed. (1986, reissued 1990), provides in-depth information on egg chemistry, composition, specialized processes, functional properties, quality measurements, and new uses for eggs and egg products. G.W. FRONING, "New Product Innovations from Eggs," chapter 4 in B.J.F. HUDSON (ed.), *New and Developing Sources of Food Proteins* (1994), pp. 71–94, provides information on new processing technologies and new egg products. The work by Stadelman *et al.* cited in the previous paragraph is also useful. (G.W.F.)

Dairy products. R. EARLY (ed.), *The Technology of Dairy Products* (1992); and ALAN H. VARNAM and JANE P. SUTHERLAND, *Milk and Milk Products: Technology, Chemistry, and Microbiology* (1994), treat the general field of dairy technology and provide a broad view of processing considerations for dairy products. Y.U. HUI (ed.), *Dairy Science and Technology Handbook*, 3 vol. (1993), brings together needed information on the principles and properties of dairy ingredients and on manufacturing technologies, applications, and engineering.

Works specifically covering the chemistry and microbiological disciplines are NOBLE P. WONG (ed.), *Fundamentals of Dairy Chemistry*, 3rd ed. (1988); and R.K. ROBINSON (ed.), *Dairy Microbiology*, 2nd ed., 2 vol. (1990). Testing and analytical procedures are covered by *Standard Methods for the Examination of Dairy Products*, 16th ed. (1993); and *Official Methods of Analysis of the Association of Official Analytical Chemists* (quinquennial). F.W. BODYFELT, J. TOBIAS, and G.M. TROUT, *The Sensory Evaluation of Dairy Products* (1988), is the best reference for organoleptic properties.

W.S. ARBUCKLE, *Ice Cream*, 4th ed. (1986), a classic text, chronicles the development of the ice cream industry, covering all aspects from manufacturing technology to techniques for dipping and serving. The most useful work for the cheese industry is FRANK KOSIKOWSKI, *Cheese and Fermented Milk Foods*, 2nd ed. (1977, reissued 1982), providing detailed explanations, descriptions, and procedures for making and enjoying cheese. P.F. FOX (ed.), *Cheese: Chemistry, Physics, and Microbiology*, 2 vol. (1987), contains more detailed scientific explanations on the cheese-making process. VINCENT L. ZEHREN and D.D. (DAVE) NUSBAUM, *Process Cheese* (1992), discusses technology. RICHARD K. ROBINSON (ed.), *A Colour Guide to Cheese and Fermented Milks* (1995); and BERNARD NANTET *et al., Cheeses of the World* (1993), are two illustrated popular texts. (D.K.B.)

Sugar.. JAMES C.P. CHEN and CHUNG-CHI CHOU, *Chen-Chou Cane Sugar Handbook: A Manual for Cane Sugar Manufacturers and Their Chemists*, 12th ed. (1993), contains detailed descriptions of raw and refined cane-sugar processes. R.A. MCGINNIS (ed.), *Beet-Sugar Technology*, 3rd ed. (1982), provides detailed and extensive descriptions of sugar beet production and beet sugar manufacture. MARGARET A. CLARKE and MARY AN GODSHALL (eds.), *Chemistry and Processing of Sugarbeet and Sugarcane* (1988), discusses recent and predicted developments in cane and beet sugar manufacture as well as the by-products of sugarcane and sugar beets. NEIL L. PENNINGTON and CHARLES W. BAKER (eds.), *Sugar: A User's Guide to Sucrose* (1990), details sugarcane and sugar beet properties and behaviour in food processing. (M.A.C.)

Cocoa products. The cocoa and chocolate industry from the growing of cocoa beans to the finished cocoa and chocolate products is covered in L. RUSSELL COOK, *Chocolate Production and Use*, 3rd ed., rev. by E.H. MEURSING (1982). JOHN SIMMONS (ed.), *Cocoa Production: Economic and Botanical Perspectives* (1976), a basic source, includes a worldwide survey of research. G.A.R. WOOD and R.A. LASS, *Cocoa*, 4th ed. (1985), discusses cocoa production in various countries. (L.R.C./Ed.)

Confectionery and candy. BERNARD W. MINIFIE, *Chocolate, Cocoa, and Confectionery: Science and Technology*, 3rd ed. (1989), deals with production methods, machinery, and formulations with scientific explanations. C. TREVOR WILLIAMS, *Chocolate and Confectionery*, 3rd ed. (1964), is a general survey of the industry with details of processes, machinery, and recipes in some sections. E. SKUSE, *Complete Confectioner*, 13th ed., rev. and edited by W.J. BUSH & CO. (1957), deals mainly with sugar confectionery. ERNEST J. CLYNE, *A Course in Confectionery*, 2 vol. in 1 (1955), also deals mainly with sugar confectionery. Journals such as *Confectionery Production* (monthly); and *The Manufacturing Confectioner* (monthly), contain useful current information. (H.B.K./B.W.M.)

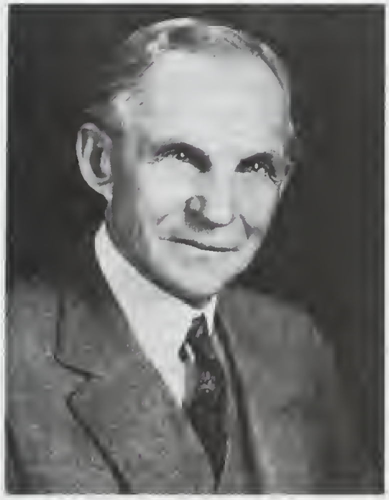
Frozen prepared foods. C.P. MALLETT (ed.), *Frozen Food Technology* (1993), is a reference book. INTERNATIONAL INSTITUTE OF REFRIGERATION, *Recommendations for the Processing and Handling of Frozen Foods*, 3rd ed. (1986), in English and French, discusses various industrial methods used in food freezing and provides recommendations on optimal conditions for freezing and for frozen storage of various food commodities. MOGENS JUL, *The Quality of Frozen Foods* (1984), describes changes in the quality of foods during frozen storage and includes calculations to determine the shelf life of frozen foods. (R.P.Si.)

Food additives. A. LARRY BRANEN, P. MICHAEL DAVIDSON, and SEPPO SALMINEN (eds.), *Food Additives* (1990), treats all aspects of the subject. THOMAS E. FURIA (ed.), *CRC Handbook of Food Additives*, 2nd ed., 2 vol. (1972, reissued 1980), contains a classification and description of the compounds used as food additives, as well as commercial forms, functions, applications, stability, and regulatory status of individual agents. JOSEPH A. MAGA and ANTHONY T. TU (eds.), *Food Additive Toxicology* (1995), discusses toxicological aspects. TIM SMITH (ed.), *Food Additive User's Handbook* (1991), provides tables containing data on the functions, characteristics, processing stability, and applications of specific food additives. (P.M.D.)

Henry Ford

Henry Ford spent most of his life making headlines, good, bad, but never indifferent. Celebrated as both a technological genius and a folk hero, Ford was the creative force behind an industry of unprecedented size and wealth that in only a few decades permanently changed the economic and social character of the United States. When young Ford left his father's farm in 1879 for Detroit, only two out of eight Americans lived in cities; when he died at age 83, the proportion was five out of eight. Once Ford realized the tremendous part he and his Model T automobile had played in bringing about this change, he wanted nothing more than to reverse it, or at least to recapture the rural values of his boyhood. Henry Ford, then, is an apt symbol of the transition from an agricultural to an industrial America.

By courtesy of the Ford Archives, Henry Ford Museum, Dearborn, Michigan



Ford, 1933.

Early life. Henry Ford was one of eight children of William and Mary Ford. He was born on the family farm near Dearborn, Michigan, then a town eight miles west of Detroit, on July 30, 1863. Abraham Lincoln was president of the 24 states of the Union, and Jefferson Davis was president of the 11 states of the Confederacy. Ford attended a one-room school for eight years when he was not helping his father with the harvest. At age 16 he walked to Detroit to find work in its machine shops. After three years, during which he came in contact with the internal-combustion engine for the first time, he returned to the farm, where he worked part-time for the Westinghouse Engine Company and in spare moments tinkered in a little machine shop he set up. Eventually he built a small "farm locomotive," a tractor that used an old mowing machine for its chassis and a homemade steam engine for power.

Ford moved back to Detroit nine years later as a married man. His wife, Clara Bryant, had grown up on a farm not far from Ford's. They were married in 1888, and on November 6, 1893, she gave birth to their only child, Edsel Bryant. A month later Ford was made chief engineer at the main Detroit Edison Company plant with responsibility for maintaining electric service in the city 24 hours a day. Because he was on call at all times, he had no regular hours and could experiment to his heart's content. He had determined several years before to build a gasoline-powered vehicle, and his first working gasoline engine was completed at the end of 1893. By 1896 he had completed his first horseless carriage, the "Quadricycle," so called because the chassis of the four-horsepower vehicle was a buggy frame mounted on four bicycle wheels. Unlike many other automotive inventors, including Charles Edgar and J. Frank Duryea, Elwood Haynes, Hiram Percy Maxim, and his Detroit acquaintance Charles Brady King, all of whom had built self-powered vehicles before Ford but who held onto their creations, Ford sold his to finance work on a second vehicle, and a third, and so on.

During the next seven years he had various backers, some of whom, in 1899, formed the Detroit Automobile Company (later the Henry Ford Company), but all eventually abandoned him in exasperation because they wanted a passenger car to put on the market while Ford insisted always on improving whatever model he was working on, saying that it was not ready yet for customers. He built several racing cars during these years, including the "999" racer driven by Barney Oldfield, and set several new speed records. In 1902 he left the Henry Ford Company, which subsequently reorganized as the Cadillac Motor Car Company. Finally, in 1903, Ford was ready to market an automobile. The Ford Motor Company was incorporated, this time with a mere $28,000 in cash put up by ordinary citizens, for Ford had, in his previous dealings with backers, antagonized the wealthiest men in Detroit.

Ford Motor Company

The company was a success from the beginning, but just five weeks after its incorporation the Association of Licensed Automobile Manufacturers threatened to put it out of business because Ford was not a licensed manufacturer. He had been denied a license by this group, which aimed at reserving for its members the profits of what was fast becoming a major industry. The basis of their power was control of a patent granted in 1895 to George Baldwin Selden, a patent lawyer of Rochester, New York. The association claimed that the patent applied to all gasoline-powered automobiles. Along with many rural Midwesterners of his generation, Ford hated industrial combinations and Eastern financial power. Moreover, Ford thought the Selden patent preposterous. All invention was a matter of evolution, he said, yet Selden claimed genesis. He was glad to fight, even though the fight pitted the puny Ford Motor Company against an industry worth millions of dollars. The gathering of evidence and actual court hearings took six years. Ford lost the original case in 1909; he appealed and won in 1911. His victory had wide implications for the industry, and the fight made Ford a popular hero.

"I will build a motor car for the great multitude," Ford proclaimed in announcing the birth of the Model T in October 1908. In the 19 years of the Model T's existence, he sold 15,500,000 of the cars in the United States, almost 1,000,000 more in Canada, and 250,000 in Great Britain, a production total amounting to half the auto output of the world. The motor age arrived owing mostly to Ford's vision of the car as the ordinary man's utility rather than as the rich man's luxury. Once only the rich had travelled freely around the country; now millions could go wherever they pleased. The Model T was the chief instrument of one of the greatest and most rapid changes in the lives of the common people in history, and it effected this change in less than two decades. Farmers were no longer isolated on remote farms. The horse disappeared so rapidly that the transfer of acreage from hay to other crops caused an agricultural revolution. The automobile became the main prop of the American economy and a stimulant to urbanization—cities spread outward, creating suburbs and housing developments—and to the building of the finest highway system in the world.

The remarkable birth rate of Model T's was made possible by the most advanced production technology yet conceived. After much experimentation by Ford and his engineers, the system that had evolved by 1913–14 in Ford's new plant in Highland Park, Michigan, was able to deliver parts, subassemblies, and assemblies (themselves built on subsidiary assembly lines) with precise timing to a constantly moving main assembly line, where a complete chassis was turned out every 93 minutes, an enormous improvement over the 728 minutes formerly required. The minute subdivision of labour and the coordination of a multitude of operations produced huge gains in productivity.

In 1914 the Ford Motor Company announced that it would henceforth pay eligible workers a minimum wage of $5 a day (compared to an average of $2.34 for the industry) and would reduce the work day from nine hours to eight, thereby converting the factory to a three-shift day. Overnight Ford became a worldwide celebrity. People either praised him as a great humanitarian or excoriated him as a mad socialist. Ford said humanitarianism had nothing to do with it. Previously profit had been based on paying wages as low as workers would take and pricing cars as high as the traffic would bear. Ford, on the other hand, stressed low pricing (the Model T cost $950 in 1908 and $290 in 1927) in order to capture the widest possible market and then met the price by volume and efficiency. Ford's success in making the automobile a basic necessity turned out to be but a prelude to a more widespread revolution. The development of mass-production techniques, which enabled the company eventually to turn out a Model T every 24 seconds; the frequent reductions in the price of the car made possible by economies of scale; and the payment of a living wage that raised workers above subsistence and made them potential customers for, among other things, automobiles—these innovations changed the very structure of society.

During its first five years the Ford Motor Company produced eight different models, and by 1908 its output was 100 cars a day. The stockholders were ecstatic; Ford was dissatisfied and looked toward turning out 1,000 a day. The stockholders seriously considered court action to stop him from using profits to expand. In 1909 Ford, who owned 58 percent of the stock, announced that he was only going to make one car in the future, the Model T. The only thing the minority stockholders could do to protect their dividends from his all-consuming imagination was to take him to court, which Horace and John Dodge did in 1916.

The Dodge brothers, who formerly had supplied chassis to Ford but were now manufacturing their own car while still holding Ford stock, sued Ford for what they claimed was his reckless expansion and for reducing prices of the company's product, thereby diverting money from stockholders' dividends. The court hearings gave Ford a chance to expound his ideas about business. In December 1917 the court ruled in favour of the Dodges; Ford, as in the Selden case, appealed, but this time he lost. In 1919 the court said that, while Ford's sentiments about his employees and customers were nice, a business is for the profit of its stockholders. Ford, irate that a court and a few shareholders, whom he likened to parasites, could interfere with the management of his company, determined to buy out all the shareholders. He had resigned as president in December 1918 in favour of his son, Edsel, and in March 1919 he announced a plan to organize a new company to build cars cheaper than the Model T. When asked what would become of the Ford Motor Company, he said, "Why I don't know exactly what will become of that; the portion of it that does not belong to me cannot be sold to me, that I know." The Dodges, somewhat inconsistently, having just taken him to court for mismanagement, vowed that he would not be allowed to leave. Ford said that if he was not master of his own company, he would start another. The ruse worked; by July 1919 Ford had bought out all seven minority stockholders. (The seven had little to complain about: in addition to being paid nearly $106,000,000 for their stock, they received a court-ordered dividend of $19,275,385 plus $1,536,749 in interest.) Ford Motor Company was reorganized under a Delaware charter in 1920 with all shares held by Ford and other family members. Never had one man controlled so completely a business enterprise so gigantic.

The winning Ford Model T entry on the road during the 1909 transcontinental race, New York City to Seattle, Washington.

By courtesy of Ford Motor Company Archives

The planning of a huge new plant at River Rouge, Michigan, had been one of the specific causes of the Dodge suit. What Ford dreamed of was not merely increased capacity but complete self-sufficiency. World War I, with its shortages and price increases, demonstrated for him the need to control raw materials; slow-moving suppliers convinced him that he should make his own parts. Wheels, tires, upholstery, and various accessories were purchased from other companies around Detroit. As Ford production increased, these smaller operations had to speed their output; most of them had to install their own assembly lines. It became impossible to coordinate production and shipment so that each product would arrive at the right place and at the right time. At first he tried accumulating large inventories to prevent delays or stoppages of the assembly line, but he soon realized that stockpiling wasted capital. Instead he took up the idea of extending movement to inventories as well as to production. He perceived that his costs in manufacturing began the moment the raw material was separated from the earth and continued until the finished product was delivered to the consumer. The plant he built in River Rouge embodied his idea of an integrated operation encompassing production, assembly, and transportation. To complete the vertical integration of his empire, he purchased a railroad, acquired control of 16 coal mines and about 700,000 (285,000 hectares) acres of timberland, built a sawmill, acquired a fleet of Great Lakes freighters to bring ore from his Lake Superior mines, and even bought a glassworks.

Vertical integration at River Rouge

The move from Highland Park to the completed River Rouge plant was accomplished in 1927. At 8 o'clock any morning, just enough ore for the day would arrive on a Ford freighter from Ford mines in Michigan and Minnesota and would be transferred by conveyor to the blast furnaces and transformed into steel with heat supplied by coal from Ford mines in Kentucky. It would continue on through the foundry molds and stamping mills and exactly 28 hours after arrival as ore would emerge as a finished automobile. Similar systems handled lumber for floorboards, rubber for tires, and so on. At the height of its success the company's holdings stretched from the iron mines of northern Michigan to the jungles of Brazil, and it operated in 33 countries around the globe. Most remarkably, not one cent had been borrowed to pay for any of it. It was all built out of profits from the Model T.

Later years. The unprecedented scale of that success, together with Ford's personal success in gaining absolute control of the firm and driving out subordinates with contrary opinions, set the stage for decline. Trusting in what he believed was an unerring instinct for the market, Ford refused to follow other automobile manufacturers in offering such innovative features as conventional gearshifts (he held out for his own planetary gear transmission), hydraulic brakes (rather than mechanical ones), six- and

The Model A and the Ford V-8

eight-cylinder engines (the Model T had a four), and choice of colour (from 1914 every Model T was painted black). When he was finally convinced that the marketplace had changed and was demanding more than a purely utilitarian vehicle, he shut down his plants for five months to retool. In December 1927 he introduced the Model A. The new model enjoyed solid but not spectacular success. Ford's stubbornness had cost him his leadership position in the industry; the Model A was outsold by General Motors' Chevrolet and Chrysler's Plymouth and was discontinued in 1931. Despite the introduction of the Ford V-8 in 1932, by 1936 Ford Motor Company was third in sales in the industry.

A similar pattern of authoritarian control and stubbornness marked Ford's attitude toward his workers. The $5 day that brought him so much attention in 1914 carried with it, for workers, the price of often overbearing paternalism. It was, moreover, no guarantee for the future; in 1929 Ford instituted a $7 day, but in 1932, as part of the fiscal stringency imposed by falling sales and the Great Depression, that was cut to $4, below prevailing industry wages. Ford freely employed company police, labour spies, and violence in a protracted effort to prevent unionization and continued to do so even after General Motors and Chrysler had come to terms with the United Automobile Workers. When the UAW finally succeeded in organizing Ford workers in 1941, he considered shutting down before he was persuaded to sign a union contract.

During the 1920s, under Edsel Ford's nominal presidency, the company diversified by acquiring the Lincoln Motor Car Company, in 1922, and venturing into aviation. At Edsel's death in 1943 Henry Ford rcsumed the presidency and, in spite of age and infirmity, held it until 1945, when he retired in favour of his grandson, Henry Ford II.

Henry Ford was a complex personality. Away from the shop floor he exhibited a variety of enthusiasms and prejudices and, from time to time, startling ignorancc. His dictum that "history is more or less bunk" was widely publicized, as was his deficiency in that field revealed during cross-examination in his million-dollar libel suit against the *Chicago Tribune* in 1919; a *Tribune* editorial had called him an "ignorant idealist" because of his opposition to U.S. involvement in World War I, and while the jury found for Ford it awarded him only six ccnts. One of Ford's most publicized acts was his chartering of an ocean liner to conduct himself and a party of pacifists to Europe in November 1915 in an attempt to cnd thc war by means of "continuous mediation." The so-called Peace Ship episode was widcly ridiculed. In 1918, with the support of Pres. Woodrow Wilson, Ford ran for a U.S. Senate seat from Michigan. He was narrowly defeated after a campaign of personal attacks by his opponent.

In 1918 Ford bought a newspaper, *The Dearborn Independent*, and in it published a series of scurrilous attacks on the "International Jew," a mythical figure he blamed for financing war; in 1927 he formally retracted his attacks and sold the paper. He gave old-fashioned dances at which capitalists, European royalty, and company executives were introduced to the polka, the Sir Roger de Coverley, the mazurka, the Virginia reel, and the quadrille; he established small village factories; he built one-room schools in which vocational training was emphasized; he experimented with soybeans for food and durable goods; he sponsored a weekly radio hour on which quaint essays were read to "plain folks"; he constructed Greenfield Village, a restored rural town; and he built what later was named the Henry Ford Museum and filled it with American artifacts and antiques from the era of his youth when American society was almost wholly agrarian. In short, he was a man who baffled even those who had the opportunity to observe him close at hand, all except James Couzens, Ford's business manager from the founding of the company until his resignation in 1915, who always said, "You cannot analyze genius and Ford is a genius."

Ford died at home on April 7, 1947, exactly 100 years after his father had left Ireland for Michigan. His holdings in Ford stock went to the Ford Foundation, which had been set up in 1936 as a means of retaining family control of the firm and which subsequently became the richest private foundation in the world.

BIBLIOGRAPHY. Ford was the nominal coauthor of three books in collaboration with SAMUEL CROWTHER: *My Life and Work* (1922, reprinted 1987), *Today and Tomorrow* (1926, reprinted 1988), and *Moving Forward* (1930). Especially recommended studies of his life and activities are ALLAN NEVINS and FRANK ERNEST HILL, *Ford: The Times, the Man, the Company* (1954), *Ford: Expansion and Challenge, 1915–1933* (1957), and *Ford: Decline and Rebirth, 1933–1962* (1963); CAROL W. GELDERMAN, *Henry Ford: The Wayward Capitalist* (1981), a full-length biography and a study of his company's development; ROBERT LACEY, *Ford: The Men and the Machine* (1986), beginning with a biography of Ford and progressing to a history of the following generations of the Ford family; ROGER BURLINGAME, *Henry Ford* (1955, reissued 1969), a short profile; WILLIAM GREENLEAF, *Monopoly on Wheels* (1961), a discussion of the Selden patent case; LOUIS P. LOCHNER, *Henry Ford: America's Don Quixote* (1925), about the Peace Ship voyage to Europe; BARBARA S. KRAFT, *The Peace Ship* (1978), a more recent treatment; and REYNOLD M. WIK, *Henry Ford and Grass-Roots America* (1972), a catalog of fan letters received by Ford. DAVID L. LEWIS, *The Public Image of Henry Ford* (1976), examines the media's portrayal of Ford and his company as well as the company's efforts to influence that portrayal. (C.W.G.)

Forestry and Wood Production

Forestry is the science of developing and managing woodlands, along with the wastelands and waters and other resources associated with them, for the benefit of humankind. The meaning of the term "benefit" in this definition may be either narrowly or broadly understood. A narrow understanding of the term might be as the maximization of the production of immediately marketable timber and the minimization of associated production costs; a broader understanding might look to some optimal sustained rate of yield consistent with the preservation of such other long-term goods as wildlife and recreational opportunities. The evolution of the practice of forestry has to a large degree paralleled that of natural-resource conservation generally. As a consequence, while the chief object of forestry is usually the raising and harvesting of successive crops of timber, professional foresters have increasingly become involved in activities related to the conservation of soil, water, and wildlife resources and to recreation.

This article deals with two distinct but closely related topics. First, it traces the history of forestry from its origin in ancient practices to its development as a scientific profession in the modern world, and it discusses the kinds and distribution of forests as well as the principal techniques and methods of modern forest management in some detail. Second, the nature of the principal forest product—wood—is discussed in the context of the ways in which it is typically processed and utilized.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 355, 724, and 731, and the *Index.*

This article is divided into the following sections:

Forestry

HISTORY

The ancient world. It is believed that *Homo erectus* used wood for fire at least 750,000 years ago. The oldest evidence of the use of wood for construction, found at the Kalambo Falls site in Tanzania, dates from some 60,000 years ago. Early organized communities were located along waterways that flowed through the arid regions of India, Pakistan, Egypt, and Mesopotamia, where scattered trees along riverbanks were used much as they are today—for fuel, construction, and handles for tools. Writers of the Old Testament make frequent reference to the use of wood. Pictures in Egyptian tombs show the use of the wooden plow and other wooden tools to prepare the land for sowing. Carpenters and shipwrights fabricated wooden boats as early as 2700 BC. Theophrastus, Varro, Pliny, Cato, and Virgil wrote extensively on the subject of trees, their classification, manner of growth, and the environmental characteristics that affect them.

The Romans took a keen interest in trees and imported tree seedlings throughout the Mediterranean region and Germany, establishing groves comparable to those in Carthage, Lebanon, and elsewhere. The fall of the Roman Empire signaled an end to conservation works throughout the Mediterranean and a renewal of unregulated cutting, fire, and grazing of sheep and goats, which resulted in the destruction of the forests. This, in turn, caused serious soil loss, silting of streams and harbours, and the conversion of forest to a scrubby brush cover known as maquis.

Medieval Europe. In medieval Europe, forest laws were aimed initially at protecting game and defining rights and responsibilities. Hunting rights were vested in the feudal lord who owned the property and who had the sole right to cut trees and export timber. Peasants were permitted to gather fuel, timber, and litter for use on their own properties and to pasture defined numbers of animals. By 1165, however, land clearing for agriculture had gone so far that Germany forbade further forest removal. The systematic management of forests had its true beginnings, however, in the German states during the 16th century. Each forest property was divided into sections for timber harvesting and regeneration to ensure a sustainable yield of timber for the entire property. This working plan called for accurate maps and assessments of timber volume and expected growth rates.

The first forest nursery

Trees have been raised from seed or cuttings since biblical times, but the earliest record of a planned forest nursery is that of William Blair, cellarer to the Abbey of Coupar Angus in Scotland, who raised trees to grow in the Highland Forest of Ferter as early as 1460. After the dissolution of the monasteries, many newly rich landowners in Scotland and England found a profitable long-term investment in artificial plantations established on poor land. John Evelyn, a courtier in the reign of Charles II, published his classic textbook *Sylva* in 1664, exhorting them to do so, and today virtually all of Britain's 2,100,000 hectares (5,200,000 acres) of woodland consists of artificial plantations. Other countries had managed their natural forests better and had little need, until recently, to afforest bare land. The 20th century, however, has seen a tremendous expansion of artificial plantations in all the continents, planned to meet the ever-growing needs for wood and paper as essential materials in modern civilization.

Modern developments. Formal education in forestry began about 1825 when private forestry schools were established. These were the outgrowth of the old master schools such as Cotta Master School, which developed into the forestry college at Tharandt—one of the leading forestry schools in Germany. The National School of Forestry was established in Nancy, Fr., in 1825.

During the 19th century the reputation of German foresters stood so high that they were employed in most continental European countries. Early American foresters, including the great conservation pioneer Gifford Pinchot, gained their training at European centres. But the doctrine of responsible control had to fight a hard battle against timber merchants who sought quick profits.

The 20th century has seen the steady growth of national forest laws and policies designed to protect woodlands as enduring assets. Beginning in the 1940s vast land reclamation was undertaken by Greece, Israel, Italy, Spain, and the Maghrib countries of North Africa to restore forests to the slopes laid bare by past abuse. The main objective of the tree planting is to save what remains of the soil and to protect the watersheds. In China, where forests once extended over 30 percent of the land, centuries of overcutting, overgrazing, and fires reduced this proportion to approximately 7 percent. China has taken major steps to improve land use, including construction of reservoirs and a huge forest planting program, which reported the planting of 15,750,000 hectares between 1950 and 1957 alone.

Forest policies around the world

The character of forest policies around the world reflects national political philosophies. In Communist countries all forests are owned by the state. In the United States both the federal and the state governments have deemed it prudent to hold substantial areas of natural forest, while allowing commercial companies and private individuals to own other areas outright. Similar patterns of ownership are found throughout most of Asia, western Europe, and the Commonwealth countries. In Japan the extensive forests are largely state owned. Tribal ownership is found in many African countries and proves a serious obstacle to effective modern management. International cooperation is effected by the Forestry Department of the United Nations' Food and Agriculture Organization, with headquarters in Rome.

Development of U.S. policies. The history of forestry in the United States followed the same path as forestry in Europe—land clearing, repeated burning, overcutting, and overgrazing—until a bill was passed by Congress in 1891 authorizing the president to set apart from the public domain reserves of forested land. In 1905 an act of Congress, with strong encouragement from President Theodore Roosevelt, transferred the Bureau of Forestry from the Department of the Interior to the Department of Agriculture. Gifford Pinchot, who had been chief of the bureau, was made chief of the newly named Forest Service. Pinchot developed the U.S. Forest Service into a federal agency that today is recognized worldwide for its research, education, and land and forest management expertise. On the state level the Morrill Act of 1862 provided for federal-state cooperative programs in which the federal government granted first land, then money, to the states for the establishment of technical agricultural colleges. The Weeks Act of 1911 authorized the federal government to assist the states in protecting forests from fire, and the Clark-McNary Act of 1924 extended the provisions of the Weeks Act to include cooperation in forest extension, planting, and assistance to forest owners.

Civilian Conservation Corps

During the Great Depression of the 1930s the interests of forestry were served most imaginatively and thoroughly by the Civilian Conservation Corps (CCC), which planted trees, fought forest fires, and improved access to woodlands across the United States. The CCC, rooted in the system of public works initiated by President Franklin D. Roosevelt, continued until 1942, acquainting many people with forestry as a major government activity.

The complete mobilization of resources for the U.S. involvement in World Wars I and II and the pent-up demand for consumer goods made heavy demands on forest resources and industries. As a result, forestry on a national basis entered a period of the most rapid advance since the turn of the century. This time the advance was stimulated by the need for forest products and by the conviction on the part of the major timber companies that they must protect their raw material supply. To protect the forests from growing pressure from single-interest groups, Congress passed the Multiple Use-Sustained Yield Act of 1960. This act directed that the national forests be managed under principles of multiple use so as to produce a sustained yield of products and services. The Bureau of Outdoor Recreation was established shortly thereafter in the Department of the Interior. The Land and Water Conservation Fund, established in 1964, launched a comprehensive program for planning and developing outdoor recreation facilities. State forestry programs had their beginnings in the United States during colonial times, but it was the Weeks and Clark-McNary laws that provided the impetus to develop recognized state forestry departments. The Smith-Lever Act of 1914 allotted funds through the state agricultural colleges for extension work in forestry. Initial programs emphasized tree planting and demonstrations, but today all aspects of forestry and natural and related resources are included.

Industrial forestry began around 1912 when Finch, Pruyn, and Company started a forestry program on its Adirondack holding in New York. Trees to be cut were marked by foresters, and the cutting budget was projected on a sustained-yield basis. A rapid expansion of company forestry programs in the northeastern United States began in the late 1920s and early 1930s. Following World War II, paper companies expanded rapidly throughout the South and West and to a lesser extent in the Northeast. Pulp and paper companies were quick to recognize the benefits to be realized from research financed by the Forest Service and by universities in such fields as tree physiology, entomology, genetics, and tree improvement. A few companies established their own experimental forests and research teams.

Citizens' organizations

The cause of forestry in the United States also has been advanced by citizens' organizations. These vary from lay and youth organizations, such as the Boy Scouts and garden clubs, to the nation's most prestigious scientific societies. The American Association for the Advancement of Science stimulated Congress in 1876 to embark on a sustained federal forestry program. The National Academy of Sciences 1896 report on forest reserves began its long involvement in forest conservation. The Society of American Foresters, founded in 1900, together with its sister societies in Canada and Mexico, represents the profession of forestry in North America.

CLASSIFICATION AND DISTRIBUTION OF FORESTS

Botanical classification places forest trees into two main groups, Gymnospermae and Angiospermae. The gymnosperms consist exclusively of trees and woody shrubs, whereas the angiosperms are a diverse group of plants that include trees and shrubs as well as grasses and herbaceous plants. The gymnosperms probably gave rise to the angiosperms, although the manner in which this took place is disputed.

Gymnosperms. The gymnosperms are of very ancient lineage and include the earliest trees on the evolutionary scale. With certain exceptions, the seeds of gymnosperms are borne in cones, where they develop naked or exposed on the upper surface of the cone scales. The wood of these trees has a simple structure. Many species are extinct, such as the tree ferns of the Carboniferous Period (280,000,000 to 345,000,000 years ago), and are known only as fossils. The ginkgo, or maidenhair tree, is the sole survivor of an entire order of gymnosperms, the Ginkgoales. Among the gymnosperms, the most important and numerous forest trees are the conifers, also known as softwoods.

This group includes the well-known pines, spruces, firs, cedars, junipers, hemlocks, and sequoias. These species are so dominant in the gymnosperm class that forests of gymnosperm trees are typically called coniferous forests. Except for the ginkgo, larches, and bald cypress, all gymnosperms are evergreen.

Angiosperms. The angiosperms constitute the dominant plant life of the present geologic era. They are the products of a long line of evolutionary development that has culminated in the highly specialized organ of reproduction known as the flower, in which seed development occurs within an ovary. This group includes a large variety of broad-leaved trees, most with a deciduous leaf habit but some that are evergreen. The angiosperms are further divided into monocots and dicots. Trees are represented in both groups.

Monocots. The monocots include principally the palms and bamboos. Palm trees form extensive savannas in certain tropical and subtropical zones but are more usually seen along watersides or in plantations.

Palm trees have no growth rings, being made up of spirally arranged bundles of fibres, giving a light, spongy wood. Palms are valuable, however, for their various fruits (coconuts, dates, and palm kernels) and leaf products (carnauba wax, raffia, and thatching and walling materials for houses in the tropics).

Bamboo

Another form of tropical monocotyledonous forest is the bamboo thicket, common in Asia, composed of giant woody grasses. One of the most versatile plants in the world, bamboo is valuable as a construction material, as well as for hundreds of other applications. Its young shoots are eaten as vegetables and are a valuable source of certain enzymes.

Dicots. Finally, a more highly evolved group of forest trees is the dicots, or broad-leaved trees, also called hardwoods. Their wood structure is complex, and each sort of broad-leaved lumber has characteristic properties that fit it for particular uses.

Occurrence and distribution. Approximately 4,000,000,000 hectares, or about one-third of the total land area in the world, is covered with closed forests of broad-leaved and coniferous species and open forests or savannas (Table 1). Because of the varying characteristics of individual tree species, the kind and distribution of the world's forests are largely determined by local conditions. Each combination of temperature, rainfall, and soil has a peculiar association of trees and other vegetation that are best equipped to compete with other plants for that site. The open forest is characteristically a tropical grassland, often disturbed by fire, with forest along streams and scattered individual trees or small groves. Closed thorn forests usually appear adjacent to the savannas. In general, coniferous forests are found in the cooler, drier areas, and the broad-leaved species are predominant in the warmer, usually moister parts of the world. Tropical forests consist almost exclusively of broad-leaved species. Mixed broad-leaved and coniferous forests are found near the boundaries between these two climatic zones.

The northern coniferous forests

Coniferous forests are largely found in the temperate climate of the Northern Hemisphere, where they cover approximately 1,100,000,000 hectares; some 85 percent of them are in North America and the erstwhile Soviet Union (Figure 1). The northern coniferous forest, or taiga, extends across North America from the Pacific to the Atlantic, across northern Europe through Scandinavia and Russia, and across Asia through Siberia to Mongolia, northern China, and northern Japan. It has outliers along all the temperate mountain ranges, including the Rockies, the Appalachians, the Alps, the Urals, and the Himalayas. Its principal trees are spruces (of the genus *Picea*), northern pines (*Pinus*), silver firs (*Abies*), Douglas firs (*Pseudotsuga*), hemlocks (*Tsuga*), and larches (*Larix*). Together these northern softwood forests form a world resource of tremendous importance, yielding the bulk of the lumber and pulpwood handled commercially. Northern conifers from many lands are extensively planted in Europe, including the British Isles.

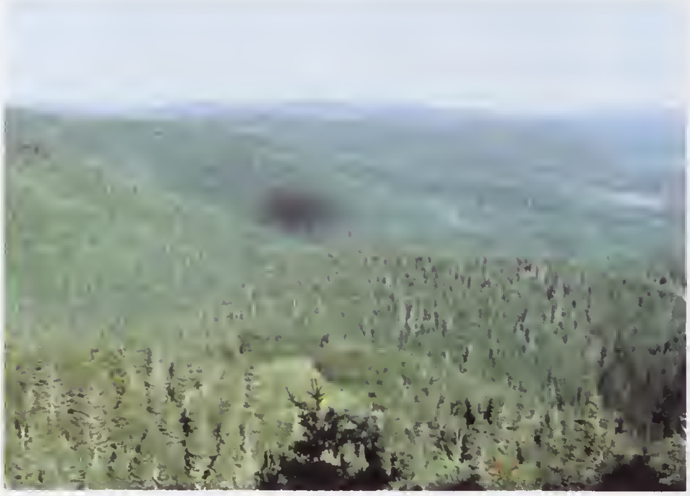


Figure 1: Northern coniferous forest, the Percé region, Gaspé Peninsula, Quebec, Can.

Francois Morneau/VALAN PHOTOS

Table 1: Distribution of the World's Forest Land*

region	total land area	closed forest		open forest	total forest area	percent of total land area forested
		broad-leaved	coniferous			
North America	1,835	168	301	215	684	37
Europe	472	65	88	21	174	37
Former Soviet Union	2,227	147	645	128	920	41
Africa	2,966	216	2	500	718	24
Latin America	2,054	666	26	250	942	46
Asia	2,573	414	55	98	567	22
Pacific area	950	50	22	70	142	15
World totals	13,077	1,726	1,139	1,282	4,147	32

*In millions of hectares.

The southern coniferous forest has a discontinuous spread through the southern part of the Northern Hemisphere, including California, the southeastern states of the United States, the Mediterranean lands of southern Europe, North Africa, Asia Minor, parts of the Asian mainland, and southern Japan. Pines are the principal trees, along with cypresses (*Cupressus* and *Chamaecyparis*), cedars (*Cedrus*), and redwoods and mammoth trees (*Sequoia* and *Sequoiadendron*). Certain southern pines such as the California Monterey pine (*Pinus radiata*) grow poorly in their native habitat but exceptionally fast when planted in subtropical Europe, Africa, New Zealand, and Australia.

In addition to the plantations of introduced pines, small areas of coniferous forest are found in the Southern Hemisphere, notably the Chile pine, *Araucaria araucana,* in the Andes; hoop pine, or bunyabunya, *Araucaria bidwillii,* in Australia; and kauri pine, *Agathis australis,* in New Zealand.

The dicotyledonous broad-leaved species form three characteristic types of forests: temperate deciduous, subtropical evergreen, and tropical evergreen.

Temperate deciduous broad-leaved forests are made up of the summer-green trees of North America, northern Europe, and the temperate regions of Asia and South America. Characteristic trees are oaks (*Quercus* species), beeches (*Fagus* and *Nothofagus*), ash trees (*Fraxinus*), birches (*Betula*), elms (*Ulmus*), alders (*Alnus*), and sweet chestnuts (*Castanea*). Temperate broad-leaved trees expand their foliage in spring, grow rapidly in summer, and shed all their leaves each fall.

Subtropical evergreen broad-leaved forests grow largely in countries with a Mediterranean type of climate—*i.e.,* hot, dry summers and cool, moist winters. Their trees have characteristic thick, hard-surfaced, leathery-textured leaves with waxy coatings that enable them to resist water loss during summer droughts. Their evergreen habit enables them to make use of moist winters. Typical trees are the evergreen oaks, species of *Quercus,* and the madrone, or *Arbutus,* while in Australia most evergreen broadleaf trees are species of *Eucalyptus.* Few evergreen broadleaf trees have high timber value, and many are little more than scrub, highly inflammable during hot, dry summers. Their world distribution embraces California; the southeastern states of the United States; Mexico; parts of Chile and Argentina; the Mediterranean shores of Europe, Asia, and North Africa; South Africa; and most of Australia.

Tropical rain forests

Tropical evergreen broad-leaved forests, or tropical rain forests, grow in the hot, humid belt of high rainfall that follows the Equator around the globe. They occur in West and Central Africa, South Asia, the northern zone of Australia, and in Central and South America. Where they extend into regions of seasonal rainfall, such as monsoon zones, they become less truly evergreen, holding many trees that stand leafless during the short dry seasons. Tropical rain forests hold a great variety of tree species. A few of the timbers, such as teak, *Tectona grandis,* in India, and mahogany, *Swietenia macrophylla,* in Central America, have uniquely useful properties or ornamental appearance and hence a high commercial value. Balsa, *Ochroma pyramidale,* from Central America, is the lightest timber known; it is used for rafts, aircraft construction, and insulation against noise, heat, and cold.

Trees outside areas classified as forestland, such as those in windbreaks, along rights-of-way, or around farm fields, are also important resources, especially in densely populated areas. For example, some 20 percent of Rwanda's farmland is maintained by farmers as woodlots and wooded pastures. These roughly 200,000 hectares of dispersed trees exceed the combined area of the country's natural forests and state and communal plantations. In the Kakamega District of Kenya more than 90 percent of the farms have scattered trees maintained for animal fodder and fuelwood. Of the 7,200,000,000 trees planted in the densely settled plains region of China, 5,800,000,000 have been planted around homes and in villages, with each household tending an average of 74 trees. Even in France, where trees are not used much for fuelwood, trees outside the forests occupy 883,000 hectares. There are no good estimates of the worldwide totals of such scattered trees, but their existence provides many locally useful products and extends the resources in the forested areas.

PURPOSES AND TECHNIQUES OF FOREST MANAGEMENT

Multiple-use concept. The forests of the world provide numerous amenities in addition to being a source of wood products. The various public, industrial, and private owners of forestland may have quite different objectives for the forest resources they control. Industrial and private owners may be most interested in producing a harvestable product for a processing mill. However, they also may want other benefits, such as forage for grazing animals, watershed protection, recreational use, and wildlife habitat. On public lands the multiple-use land management concept has become the guiding principle for enlightened foresters. This is a complex ecological and sociological concept in contrast to the single-use principle of the past. The challenge, in the words of Gifford Pinchot, is to "ensure the greatest good for the most people over the long run." Thus timber production may have top priority in certain areas, but in others, such as those near large population centres, recreational values, for example, may have high priority. Multiple use calls for exceptional skill on the part of forest managers.

Sustained yield. Forest management originated in the desire of the large central European landowners to secure dependable income to maintain their castles and retinues of servants. Today forest management is still primarily economic in essence, because modern forest industries, mainly sawmilling and paper manufacture, can be efficient only on a continuous-operation basis.

Rotation periods

Foresters think in long time scales, in line with the long life of their renewable crop. However, it is possible that a forest can be managed in such a way that a modest timber crop may be harvested indefinitely year after year if annual harvest and the losses due to fire, insects, diseases, and other destructive agents are counterbalanced by annual growth. This is the sustained-yield concept. An important element is the rotation, or age to which each crop can be grown before it is succeeded by the next one. Examples of short rotation periods in the subtropics are seven years for leucaena for fuelwood, 10 years for eucalyptus, and 20 years for pine for pulpwood. Here a sustained yield could in theory be obtained simply by felling one-tenth of the eucalyptus forest each year and replanting it. Rotation periods for pulpwood in northern Europe and North America extend to 50 years. Softwood sawlogs often need 100 years to reach an economic size, while rotation periods for broad-leaved trees, such as oak and beech, in central Europe, may extend to two centuries. Over so long a growing spell only part of the lumber yield is obtained by the clear-cutting of a small fraction of the forest each year. The rest is secured by systematically thinning out the whole forest periodically.

Sustained-yield principles are likewise applied to minor forest produce. Turpentine and pitch, also known as naval stores, are obtained by the systematic tapping of the lower trunk of certain subtropical pines. Successive cuts with a chisellike tool every few days during a succession of summers eventually kill the trees. To ensure continued yields, crops of young pines are raised rotationally to replace those felled. A similar system is followed for Para rubber, *Hevea brasiliensis,* grown in plantations.

Forest products. The culture of trees in natural forests and plantations for the yield of lumber, pulp, chips, and specialty products is a principal management objective. In many parts of the world the harvest of wood for firewood and charcoal is the dominant use, and these products are often in short supply. Timber stands must be felled and regenerated in an orderly sequence to meet continuing industrial demands.

Silvics

Silviculture. Silviculture is the branch of forestry concerned with the theory and practice of controlling forest establishment, composition, and growth. Like forestry itself, silviculture is an applied science that rests ultimately upon the more fundamental natural and social sciences. The immediate foundation of silviculture in the natural sciences is the field of silvics, which deals with the laws underlying the growth and development of single trees and of the forest as a biologic unit. Growth, in turn, depends on local soils and climate, competition from other vegetation, and interrelations with animals, insects, and other organisms, both beneficial and destructive. The efficient practice of silviculture demands knowledge of such fields as ecology, plant physiology, entomology, and soil science and is concerned with the economic as well as the biologic aspects of forestry. The implicit objective of forestry is to make the forest useful to man.

The practice of silviculture is divided into three areas: methods of reproduction, intermediate cuttings, and protection. In every forest the time comes when it is desirable to harvest a portion of the timber and to replace the trees removed with others of a new generation. The act of replacing old trees, either naturally or artificially, is called regeneration or reproduction, and these two terms also refer to the new growth that develops. The period of regeneration begins when preparatory measures are initiated and does not end until young trees have become established in acceptable numbers and are fully adjusted to the new environment. The rotation is the period during which a single crop or generation is allowed to grow.

Intermediate cuttings are various types of cuttings made during the development of the forest—*i.e.,* from the reproduction stage to maturity. These cuttings or thinnings are made to improve the existing stand of trees, to regulate growth, and to provide early financial returns, without any effort directed at regeneration. Intermediate cuttings are aimed primarily at controlling growth through adjustments in stand density, the regulation of species composition, and selection of individuals that will make up the harvest trees. Protection of the stand against fire, insects, fungi, animals, and atmospheric disturbances is as much a part of silviculture as is harvesting, regenerating, and tending the forest crop.

Silvicultural systems are divided into those employing natural regeneration, whereby tree crops are renewed by natural seeding or occasionally sprout regrowth, and those involving artificial regeneration, whereby trees are raised from seed or cuttings. Natural regeneration is easier but may be slow and irregular; it can only renew existing forests with the same sorts of tree that grew before. Artificial regeneration needs more effort, yet can prove quicker, more even, and in the long run more economical. It permits the introduction of new sorts of trees or better strains of the preexisting ones.

Natural regeneration. In established forests the selective cutting of marketable timber, taking either one tree at a time (single-tree selection) or a number of trees in a cluster (group selection) and leaving gaps in which replacements can grow up from natural seedlings, can prove economical and also ensure the best possible use of available soil, light, and growing space. The best examples of single-tree-selection forests are found in Switzerland, on slopes where any clear felling could lead quickly to soil erosion and avalanches.

Alternative methods

Alternative methods of natural regeneration deal with areas of land as units, rather than with single trees. One highly effective example is employed in the Douglas fir forests along the Pacific slope of Canada and the western United States. Logging by powerful yarding machines, using overhead cables, creates wedge-shaped gaps of cleared ground. The surrounding forest is left standing for many years in order to provide shelter and seed. Abundant seed is carried by wind on to the cleared land and gives rise, in a few years, to a full crop of seedling firs. After these have reached seed-bearing age, the areas previously left standing may be removed in their turn. Similar systems using a pattern of strips cut across the forest, or circular plots gradually extended until they meet and coalesce, are employed in France and Germany.

A silvicultural system employing practices of short rotation (five to 10 years) and intensive culture (fertilization, weed, and insect and disease control) with superior genotypes relies on coppice, or regeneration from sprouts arising from stumps of felled trees, as the method of regeneration of the new crop and is characterized by high productivity.

Artificial regeneration. Artificial regeneration is accomplished by the planting of seedlings (the most common method) or by the direct planting of seeds. Direct seeding is reserved for remote or inaccessible areas where seedling planting is not cost-effective. A few tree species, such as poplars (*Populus* species) and willows (*Salix* species), are artificially reproduced from cuttings. Most forest planting in North America involves the conifers, especially the pines, spruces, and Douglas fir, because of the prospects of successful establishment and high financial yield. The amount of hardwood planted worldwide has increased from earlier periods, with major gains in tropical hardwoods (*Eucalyptus* species, *Gmelina* species) and high value temperate species.

Artificial regeneration offers greater opportunity than natural regeneration to modify the genetic constitution of stands. The most important decision made in artificial regeneration is the selection of the species used in each new stand. The species chosen should be adapted to the site. The most successful introductions are obtained by moving species to the same latitude and position on the continent that they occupied in their native habitat. For example, many conifers of the western coasts of North America have been successful at the same latitudes in western Europe. The forest economy of many countries in the Southern Hemisphere is dependent on pines introduced from localities of comparable climate in the southern United States, California, and Mexico.

Seed orchards

The variability of seed quantity and quality and the demand for superior genotypes has led to the creation of seed orchards, stands of trees selected for superior genetic characteristics, which are cultivated to produce large quantities of seed. Most kinds of seed can be stored in sealed containers in refrigerators at temperatures near freezing for several years without a significant loss in viability. For some species, a brief period of cold storage may be necessary for the seeds to germinate; this stratification treatment is needed to satisfy the dormancy requirement of some temperate-zone species.

Direct sowing of harvested seed in the forest or on open land is not a common practice because of forest seed-eaters (mice, squirrels, birds) and the problem of weed growth. Tree seedlings are therefore raised in forest nurseries, where effective protection is possible. These seedlings almost invariably come from seed, although vegetative propagation from rooted cuttings is a useful technique of perpetuating valuable strains of certain species. Seedlings grown in raised seedbeds are removed from the nursery soil when large enough and are bare-rooted when planted in the field. Seedlings grown in individual containers have an intact root system encapsulated in a soil plug for planting. In either case, the system can be highly mechanized. To enhance seedling quality, the seedbeds or container media are inoculated with specific microorganisms that form symbiotic relationships with the seedlings. These microorganisms include certain fungi, which form mycorrhizae with the roots and improve nutrient and water uptake, and nitrogen-fixing organisms such as *Rhizobium* species and *Frankia* species, which contribute nutrients. Selective herbicides, insecticides, and fungicides are applied before or after seedling emergence to keep the developing seedlings free of weeds, insects, and disease.

Many tree seedlings are suitable for field planting after a few months in a containerized seedling nursery or after one to two years in a seedbed. Slow-growing species are transplanted by hand or machine during the dormant season to transplant beds where they are root-pruned and fertilized to stimulate top growth and the development of a bushy root system, characteristics essential for survival in field planting. The mechanized operation is highly efficient. One machine and four workers can transplant 30,000–40,000 seedlings each working day. Weeds are controlled during the transplant stage by chemical herbicides that inhibit weed seed germination or growth or by mechanical harrows drawn between the rows.

Seedling preparation

In preparation for field planting, dormant nursery-grown seedlings are undercut with a sharpened steel blade and removed from the bed by hand or by a mechanized vibrating lifter and conveyor belt system. Roots of seedlings lifted in autumn are packed with moistened sphagnum moss, and the bundles are stored in refrigerated coolers. Alternatively, seedlings may be placed in a trench, or heeling-in bed, and covered with soil and mulch until spring. At the time of lifting, seedlings should be culled to eliminate those that will not survive after planting—*i.e.,* seedlings infested with insects or disease, badly damaged in lifting and handling, having distinctly poor root systems, or falling below minimum size standards. It is imperative that the seedlings be kept cool and the root systems moist in all phases of the lifting, storage, transport, and planting processes.

Container-grown seedlings are culled in a manner similar to the bare-rooted stock and in most cases are shipped in the containers in which they were produced. The container method, which has traditionally been used in the tropics or in locations that are hot and dry, has become the principal method of seedling production in Canada, Scandinavia, and portions of continental Europe, Japan, and China.

Planting tree seedlings is one of the most costly investments in the production of a forest crop. The success of a whole rotation is often determined by the soundness of decisions made about planting. These decisions concern the selection of the planting stock, the density of the planting, the use of mixed plantings, the season of planting, preparation of the site prior to planting, and even the method of planting. In temperate climates planting is generally conducted from late winter to late spring, but the use of container-grown seedlings extends the planting season into the early summer and includes a period in early autumn.

On level ground, machine planting is preferred over hand planting. A planting machine forms a groove in the soil in which seedlings are placed at specified intervals; a set of blades then cuts into the soil around the planted seedling, and a set of packing wheels firms the soil around it. A planting machine pulled behind a single tractor on prepared level ground can set 8,000–10,000 seedlings per day. On steep slopes, broken or rocky ground, or amid tree stumps and tops, planting is done by hand. The planter uses a spade, planting bar, or mattock (or a variation of one of these) to cut a notch, or dig a pit, into which the seedling roots are inserted. Soil is then replaced and stamped firmly around the base of the seedling.

Weed control

During the following growing season, and possibly two to three years thereafter, weed control may be essential for the survival and early growth of the planted seedling. Weeds may be removed by hand with a sharp tool or hoe or by other mechanical means such as mowing or cultivating between the planted rows. Herbicides may offer a more effective and efficient means of weed control. While care must be exercised to shield the tree from many chemicals, compounds are available that kill unwanted vegetation but do not harm the tree seedling. In some regions the lower branches of conifers and certain highly valued hardwoods are pruned from saplings and young trees to improve the quality and value of the main stem and improve access into the plantation. Otherwise, the artificially established plantation needs, and receives, no more attention than does the naturally regenerated crop.

Until the 20th century foresters usually accepted the land much as they found it. Their reaction to infertile soil was to plant aggressive species of trees, regardless of their potential market value, and to accept lower returns in plant production. Development of modern machines and a growing understanding of plant nutrition and soil chemistry now enable foresters to improve sites much as a farmer does and thereby to increase output substantially. Mechanical draining, using tractor-drawn plows to create deep open drains and so aerate the soil, is now usual on the peaty swamps of Europe, especially in Finland. On the hard heathlands of Great Britain, 120,000 hectares of new afforestation land were broken up after 1940 with sturdy plows designed to turn over firmly compacted soil layers. Plowing facilitates penetration of air, water, and tree roots, checks weed growth, and lessens fire hazard. So far it has usually been confined to strips for each row of trees, but full plowing as done on a farm promises further advantages.

In the poorly drained Great Lakes states and in coastal areas in the southeastern and southern United States, sites are prepared by a bedding plow, which creates an alternative ridge and valley surface that improves soil drainage, aeration, and nutrient availability. Subsequent to bedding, seedlings are planted on the ridge or bed. Because forest crops are rarely irrigated (returns are too low for the capital cost invested), forest plantings on droughty sites require a careful selection of the species and the time for planting and an effective weed control program.

Tree nutrition

The fundamental relationship between mineral nutrition and growth is the same for trees as for other plants. An understanding of forest tree nutrition requires recognition of factors distinctive to forests: (1) The nutrient demands of the plantation vary from season to season and with the developmental stage of the stand. During the life of a forest tree crop, large quantities of nutrients are returned to the soil in organic matter, which is, in turn, mineralized and made available for reuse by the same or the following crop. (2) Retranslocation of absorbed nutrients is highly developed in trees; *i.e.*, nutrients in leaves move back into stems prior to fall leaf drop and then move into new leaves in the spring. (3) Except for the first year after planting, trees start the growing season with a developed framework for photosynthesis and an established root system for nutrient and water uptake. (4) The use of soil resources such as water and nutrients by trees may often be strongly influenced by mechanisms involved in adaptations for survival from one season to another, rather than in growth.

Judicious management of nutrition ensures not only increased productivity of existing forests but also sustained productivity over many rotations. In southern Australia, for example, declines in yield of 25–30 percent in second rotation radiata pine (*Pinus radiata*) plantations have been corrected by a number of means, including intensive silviculture (site preparation, weed control, fertilization) during the early stages, retention and management of forest debris (leaves, branches, etc.) to conserve nutrients, and intercropping with annual legumes, which supply nitrogen and other nutrients.

Range and forage. Important among the broad spectrum of forest resources are the understory plants that can provide forage for grazing animals, both domestic and wild. Grazing livestock are useful to the forest manager. Dense old-growth forest or vigorous second-growth stands with closed canopies generally have sparse, low-quality forage. Large forest management units, however, generally contain extensive logged or burned areas where understory forage plants temporarily dominate the site. These areas are transitory since the tree canopies close in 10 to 20 years, but they can provide good forage until canopy closure. Cutting cycles in the managed forest and even wildfires provide a continuing grazing resource that shifts from one location to another. In addition, open meadows occurring in valley bottoms, open forests on shallow soils, and grassland balds on windswept ridge tops greatly enrich the grazing potential of the forest. Grazing fees offset the long-term investments that must be carried in renewing the forest.

Hardwood forests are more susceptible than coniferous forests to grazing damage. The current year's growth on broad-leaved trees provides palatable forage during most seasons of the year, whereas coniferous needles are much less palatable. Uncontrolled livestock-grazing in some parts of the world has been particularly devastating to forests and is a serious problem.

Recreation activities

Recreation and wildlife. From the earliest times human beings have looked to the forests for recreation. Today, recreation in forests assumes ever-growing importance with the growth of cities whose inhabitants need a change of scene, fresh air, and freedom to wander, as a relief to the stresses of industrial and commercial life. Imaginative planning is essential to ensure that people actually find what they are seeking without damage to the forest environment or conflict with the pleasures of others. The most popular outdoor recreation activities utilize forestland and include hunting and fishing, picnicking and camping, hiking, mountain climbing, driving for pleasure, boating and other water sports, winter sports, photography, and nature study. The challenge is to balance the varied demands for recreational use with the other forest uses.

For many recreationists the main attraction of the woods is the abundance of animal and plant life. The forest manager must attempt to satisfy the diverse needs of hunters and sportsmen, outdoorsmen, and preservationists. This requires a broad expertise drawing on principles from the social sciences, natural history, wildlife management, landscape design, law, and public administration, among other disciplines.

Recreation management includes visitor management as well as resource management. Reasonably accurate assessments of the type and amount of use that areas receive are important to allow for efficient allocation of budgets and employee time and to ensure that the degree of use does not cause excessive impacts on resources and thus destroy the recreational value of the site. Skillful location of roads, picnic points, parking lots, and campgrounds ensures that the great majority of visitors congregate in relatively small portions of a large forest. Visitor management for some situations can be aided by use of computer-generated simulation models.

Some types of recreation require intensive management and special amenities. Vehicular camping facilities, for example, are designed for intensive use by large numbers of people and typically provide electrical hookups, toilets, showers, picnic tables, fireplaces, garbage receptacles, directional and interpretive signs, and play areas. These features must be durable and easily maintained. Downhill ski areas are most popular when well equipped with various runs, lifts, restaurants, and lodges. Other types of recreation, such as trail hiking and cross-country skiing, demand larger tracts of land but fewer improvements. Wilderness areas afford the personal challenge and serenity of backpacking, tent camping, and canoeing.

The interpretation of what visitors see in the forests has become a growing activity of most forest services. Nature trails, guidebooks, signposts, interpretive museums, and information stations assist visitors who come to learn as well as to enjoy.

Forests contain natural habitats for a wide range of wildlife, from the elks, wolves, lynxes, and bears of northern coniferous forests to the antelopes, giraffes, elephants, lions, and tigers of tropical savannas and jungles. Certain birds, such as pheasants, wood grouse, and quail, have high sporting value, while others are cherished for attractive song, appearance, or rarity. Many endangered species depend on forest habitats that are carefully protected by national and international laws.

Forest managers must attend to the interrelated, and sometimes directly opposed, wildlife interests of hunters, conservationists, and farmers. Obviously the same animal can present a different aspect to each group. A Bengal tiger, for example, provides a biologist with a classic example of a carnivorous beast living in harmony with a jungle environment and restraining its main prey, deer, from undue increase in numbers. But to a village peasant it is a menace to his cows and goats and a threat to the safety of himself and his family, while a game hunter regards it as a magnificent quarry demanding all his skill. The needs of the forest itself require the numbers of grazing and browsing animals to be kept to a tolerable level. Otherwise renewal of tree crops becomes impossible.

Wildlife requirements

Virtually every change that occurs in a forest benefits some wildlife species and harms others. Some species require a diversity of conditions; one type for feeding, another for nesting, and yet another for cover. Some have very specific requirements essential to their existence, whereas others have a broad range of tolerance. In any case, the life history characteristics of the species must be known in order for the resource manager to plan and implement practices necessary for the well-being of the species. Sometimes the best management involves increasing the forest edge habitat, frequented by many kinds of wildlife. Forest edge improvement may be integrated with timber harvesting and the construction of fire lanes and logging roads. Because food and cover for wildlife are often more plentiful in the early stages of forest development, retardation of succession by prescribed burning may be beneficial to wildlife. Food crops may be planted in certain areas to improve the wildlife-carrying capacity. Adjustments are often made by foresters in cutting procedures, rotation age, regeneration methods, and other practices to accommodate the food and cover needs of wildlife and fish. Certain areas may be managed exclusively for wildlife, particularly in situations where habitat for endangered species must be protected.

In virtually every country the sporting aspect of woodland wildlife management is controlled, to some degree, by general game laws, which also apply outside the forests. These prescribe licenses for firearms and the taking of specified birds and beasts; they usually lay down closed seasons during which certain game may not be shot and also set limits to the sportsman's bag of rare species. In the United States a peculiar situation exists whereby the game legislation of the separate states applies unchanged over most publicly owned forests. In other countries the forest managers are in a stronger position, since local game laws are adjusted to their particular requirements.

Watershed management and erosion control. Not only is the presence of water in soils essential to the growth of forests, but improved water yield and quality are becoming increasingly important management objectives on many forested lands. Forests and their associated soils and litter layers are excellent filters as well as sponges, and water that passes through this system is relatively pure. Forest disturbances of various kinds can speed up the movement of water from the system and, in effect, reduce the filtering action. While disturbances are inevitable, in most instances they need not contribute to poor water quality.

In mountainous territory the value of forests for watershed and erosion protection commonly exceeds their value as sources of lumber or places of recreation. The classic example is found in Switzerland and the neighbouring Alpine regions where the existence of pastoral settlements in the valley is wholly dependent on the maintenance of continuous forest cover on the foothills of the great peaks. This is combined skillfully with limited lumbering and widespread recreational use by tourists.

The guiding principle of management where erosion threatens is therefore the maintenance of continual cover. Ideally, this is achieved by single-stem harvesting; only one tree is felled at any one point, and the small gap so created is soon closed by the outward growth of its neighbours.

The progress of water, from the time of precipitation until it is returned to the atmosphere and is again ready to be precipitated, is called the hydrologic cycle. The properties of the soil plant system provide mechanisms that regulate interception, flow, and storage of water in the cycle. The water that moves downward into the soil, or infiltrates, is the difference between precipitation and the losses due to canopy interception, forest floor interception, and runoff. The amount of water stored in the soil is largely dependent on the physical properties of the soil, its depth, and the amount of water lost due to evaporation from the soil surface and transpiration from plants (evapotranspiration). Transpiration is the water absorbed by plant roots that is subsequently evaporated from their leaf surfaces. Deep forest soils have a high water-storage capacity. Unless they are very porous and drain freely, they have a water table below which the subsoil is saturated. The depth of the water table varies seasonally and is higher during periods of low evapotranspiration. Removal of the forest canopy in wet areas also raises the water table. Most tree roots need air to survive and cannot exploit soil below the water table. The drainage of land having a high water table usually increases the productivity of the forest.

Transpiration

When incoming precipitation exceeds the soil's water-storage capacity, the excess water flows from the soil and can be measured as streamflow. The water yield of a forest is a measure of the balance between incoming precipitation and outflow of water as streamflow. The amount of increase in water yield depends on annual precipitation as well as the type and amount of overstory vegetation removed. As forests regrow following cutting, increases in streamflow decline as a result of increased transpirational losses. Streamflow declines are greater in areas that are restocked with conifers than in those restocked with hardwoods. This results from greater transpiration losses during the winter months from coniferous species.

Despite the uncertain balance of water gain and loss, forests offer the most desirable cover for water management strategies. Water yields are gradual, reliable, and uniform, as contrasted to the rapid flows of short duration characteristic of sparsely vegetated land. Unforested land sheds water swiftly, causing sudden rises in the rivers below. Over a large river system, such as that of the Mississippi, forests are a definite advantage since they lessen the risk of floods. They also provide conditions more favourable to fishing and navigation than does unforested land. All natural streams contain varying amounts of dissolved and suspended matter, although streams issuing from undisturbed watersheds are ordinarily of high quality. Waters from forested areas are not only low in foreign substances, but they also are relatively high in oxygen and low in temperature. Nonetheless, some deterioration of stream quality can be noted during and immediately after clearcut harvesting, even under the best logging conditions. The potential for water-quality degradation following timber harvest may involve turbidity (suspended solids) as well as increases in temperature and nutrient content. Sediment arising from logging roads is the major water-quality problem related to forest activities in many areas.

The belief that forests increase rainfall has not been substantiated by scientific inquiry. Local effects can, however, prove substantial, particularly in semiarid regions where every millimetre of rain counts. The air above a forest, as contrasted with grassland, remains relatively cool and humid on hot days, so that showers are more frequent. Fog belts, such as those found along the Pacific seaboard of North America and around the peaks of the Canary Islands, give significant water yields through the interception of water vapour by tree foliage. The vapour condenses and falls in a process described as fog drip.

Fire prevention and control. A forest fire is unenclosed and freely spreading combustion that consumes the natural fuels of a forest; *i.e.,* duff, grass, weeds, brush, and trees. Forest fires occur in three principal forms, the distinctions

Types of forest fires

depending essentially on their mode of spread and their position in relation to the ground surface. Surface fires burn surface litter, other loose debris of the forest floor, and small vegetation; a surface fire may, and often does, burn taller vegetation and tree crowns as it progresses. Crown fires advance through the tops of trees or shrubs more or less independently of the surface fire and are the fastest spreading of all forest fires. Ground fires consume the organic material beneath the surface litter of the forest floor; ground fires are the least spectacular and the slowest-moving, but they are often the most destructive of all forest fires and also the most difficult to control.

A forest fire does a number of specific things. First, and perhaps most obviously, it consumes woody material. Second, the heat it creates may kill vegetation and animal life. In most fires, much more is killed, injured, or changed through heat than is consumed by fire. Third, it produces residual mineral products that may cause chemical effects, mostly in relation to the soil. The lethal temperatures for the living tissues of a tree (*i.e.,* the phloem and cambium, which are located under the bark) begin at 49° C (120° F) if exposure is prolonged for one hour. At 64° C (147° F) death is almost instantaneous. The ignition temperature for woody material is approximately 343° C (650° F), with a flame temperature of 870°–980° C (1600°–1800° F).

Forest fires seldom occur in tropical rain forests or in the deciduous broad-leaved forests of the temperate zones. But all coniferous forests, and the evergreen broadleaf trees of hot, dry zones, frequently develop conditions ideally suited to the spread of fire through standing trees. For this, both the air and the fuel must be dry, and the fuel must form an open matrix through which air, smoke, and the gases arising from combustion can quickly pass. Hot, sunny days with low air humidity and steady or strong breezes favour rapid fire spread. In coniferous forests the resinous needles, both living and dead, and fallen branch wood make an ideal fuel bed. The leaves of evergreen broadleaf trees, such as hollies, madrone, evergreen oaks, and eucalyptus, are coated in inflammable wax and blaze fiercely even when green. Once started, fire may travel at speeds of up to 15 kilometres (10 miles) per hour downwind, spreading slowly outward in other directions, until the weather changes or the fuel runs out.

Fire danger rating

Well over 95 percent of all forest fires are caused by people, while lightning strikes are responsible for 1–2 percent. In some countries the setting of fires for clearing cropland is an integral technique of agriculture. In other areas forest fire prevention, including public education, hazard reduction, and law enforcement, consumes a considerable amount of time and money. The two basic steps in preventing forest fires are reducing risk and reducing hazard. Risk is the chance of a fire's starting as determined by the presence of activity of causal agents, most likely human beings. Hazard is reduced by compartmentalizing a forest with firebreaks (alleyways in which all vegetation is removed) and reducing the buildup of fuel (litter, branches, fallen trees, etc.) by controlled burning. In the United States the Forest Service devised a National Fire-Danger Rating System, which is the resultant of both constant and variable fire danger factors that affect the inception, spread, and difficulty of control of fires and the damage they cause.

Effective fire control begins with a field survey and map to identify the areas at risk, delineate them, and define and improve the barriers or firebreaks that may limit fire spread. Natural barriers include rivers, lakes, ridge tops, and tracts of bare land. Artificial barriers can be roads, railways, canals, and power-line tracks, but usually extra firebreaks must be cut to link these and provide wider gaps that fire cannot readily jump. Belts of land from 10 to 20 metres wide are cut clear of trees or left unplanted when a new forest is formed. Sometimes the soil is left bare and cultivated only at intervals to check invasion by weeds. Usually it is sown with an even crop of low perennial grasses or clovers and kept short by mowing or grazing. This checks soil erosion, provides an evergreen fireproof surface, and allows access on foot, by car, or in an emergency by fire-fighting trucks. Surfaced roads, serving also for lumber haulage and access for recreation, are of critical importance in fire fighting. Signposts are needed to guide fire crews unfamiliar with the woods and to mark water supplies and rendezvous points.

Detection is the first step in fire suppression. Many countries have organizations of trained professionals to detect and fight fires; others rely on volunteers or a combination of the two. Tower lookouts are the mainstay of nearly all detection systems, although the use of aircraft and satellites has modified this view in countries with an advanced fire control program. Fire surveillance is essential during seasons of high risk. Towers are set on hilltops where observers equipped with binoculars, maps, and a direction scale determine the compass direction of smoke and notify the fire control base via telephone or radio. If a fire can be seen from two or more towers, its precise position is quickly determined by mapping the intersection of cross bearings. Aircraft are used to detect fires and to carry out reconnaissance of known fires. Aerial surveillance has probably been most successful in detecting lightning-caused fires and is most often employed in areas of relatively low-value lands and inaccessible areas. An aircraft is essentially a moving fire tower, and the problems of detection that apply to a tower also apply to an aircraft; however, new developments in remote-control television, high-resolution photography, heat-sensing devices, film, and radar make fire detection by aircraft and satellite more efficient and location more accurate. Satellites provide a rapid means of collecting and communicating highly precise information in fire detection, location, and appraisal.

The "fire triangle"

Once a fire has been detected, the next step is fire suppression. The first job is to stop or slow the rate of spread of the fire, and the second job is to put it out. The aim of suppression is to minimize damage at a reasonable cost. This does not necessarily mean the same thing as minimizing the area burned, but it is a major goal. Suppression is accomplished by breaking the "fire triangle" of fuel, temperature, and oxygen by robbing the fire of its fuel (by physically removing the combustible material or by making it less flammable through application of dirt, water, or chemicals); by reducing its temperature (through application of dirt, water, or chemicals and partial removal or separation of fuels); and by reducing the available oxygen (by smothering fuels with dirt, water, fog, or chemical substances).

The great majority of all forest fires are contained by professional fire fighters equipped with numerous hand tools (spades, beaters, axes, rakes, power saws, and backpack water pumps). Trained fire crews with light, hand equipment can be carried quickly to a fire by truck, delivered by helicopter, or even dropped by parachute. When necessary, large machines (bulldozers or plows) are used to clear openings, or firebreaks, which stop the spread of the fire (Figure 2). This requires clearing surface and sometimes aerial fuels from a strip of land and then digging down to mineral soil to stop a creeping or surface fire. A control line can also be established by directly extinguishing the fire along the edge or by making fuels nonflammable. In some cases a backfire may be deliberately set between the control line and the oncoming fire to burn out or reduce the fuel supply before the main fire, or head fire, reaches the control line.

Water is the most obvious, efficient, and universal fire extinguisher, but large-scale use of water in fire fighting is limited because it is usually in short supply and application methods are not adequate. For these reasons other materials have been tested for persistence and efficiency in putting out fires. Wetting agents change the physical characteristics of water to increase its penetrating and spreading abilities. Retardants, such as sodium calcium borate, reduce the flammability of wood and therefore its rate of burning. Foaming agents in powder or liquid form can greatly increase the mixture volume and thereby cool, moisten, and insulate the fuel.

Aircraft can quickly carry in water and other chemicals to be dropped or sprayed on the fire. A method developed on Canadian lakes is to fill the floats of a seaplane with water, which is done as it skims the lake on takeoff, and to discharge this through nozzles over the fire.

The prescribed use of fire in forestland management

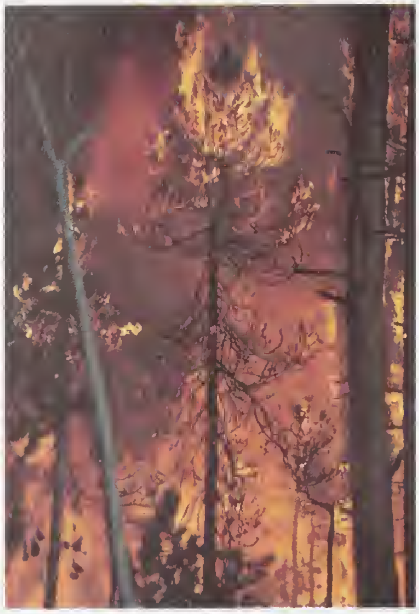


Figure 2: Forest fire in Yellowstone National Park, 1988, one of the most destructive forest fires in the history of the United States.

© Aiuppy Photographs/Laurance Aiuppy

is approached with understandable reluctance by many foresters and wildland managers. Yet, fire has a place in the management of particular ecosystems. The decision to use fire is usually based on a balancing of pros and cons; *i.e.*, damage, possible or expected, must be weighed against benefits. Under proper circumstances, prescribed burning can be used to prepare seedbeds for natural germination of most tree species, to control insect and disease infestations, to reduce weed competition, to reduce fire hazard, and to manipulate forest cover type.

Insect and disease control. Enormous numbers and varieties of insects, fungi, bacteria, and viruses occur in forests and are adapted to live on or around trees. Many of these are beneficial, and even the destructive ones are usually held in check by their natural enemies or an unfavourable environment. The normal population levels of pest organisms result in limited reduction in tree growth or the total destruction of only a small number of trees in the forest. The losses are generally accepted by foresters as unavoidable and are tolerated as long as the annual destruction does not seriously affect the net annual increase in wood production.

Every part of a growing tree—root, trunk, bark, leaf, flowers, and seeds—is potentially subject throughout every stage of its life to attack by some harmful insect or fungus (Figure 3). Insects actually destroy more standing timber than does any other agent. Bark beetles, including species of *Dendroctonus* and *Ips*, are among the most destructive insects. They bore into the tree and feed just below the bark, where they create tiny channels that disrupt the flow of food to the roots, often killing the tree. Diseases frequently retard growth of trees and are less of a factor in mortality. A particularly destructive disease is caused by fungi that decay the wood of trees. The heartrot fungi gain entrance through any wound resulting from fire scars, broken limbs, or anything else that damages the tree's protective tissue. Were it not for heartrot, a large number of conifers and broad-leaved trees could be left to grow for many more years.

Bark beetles

Insect and disease organisms accidentally introduced to forests from other parts of the world often develop serious epidemic conditions because of the lack of any natural control. Because of rapid global transportation, insects and fungal spores can be spread easily throughout the world and arrive in a healthy condition. The seriousness of the situation cannot be overestimated, and the enforcement and improvement of plant quarantine laws is essential. Typically disasters have arisen where quarantine has failed or has been imposed too late. The American chestnut, *Castanea dentata,* has been virtually wiped out by the chestnut blight fungus, *Endothia parasitica,* which does little harm to related trees in its native China. Elms have suffered severely, both in Europe and in the United States, from the elm disease fungus, *Ceratocystis ulmi,* which was first detected in The Netherlands and is carried from tree to tree by flying beetles. Minute aphids, probably introduced on living plants from Asia, now make it impossible to raise commercial crops of two conifers once valued in Britain, namely, the white pine, *Pinus strobus,* from New England, and the European silver fir, *Abies alba,* native to Switzerland.

Generally the healthier the forest, the more resistant it is to widespread pest attack. Overmature, weak, windthrown, and lightning- or fire-killed trees have little or no defense against infestation and are a factor in the buildup of pest populations. Selective cutting of susceptible trees, thinning that accelerates growth, and other similar long-range forest management practices that stimulate vigorous tree growth are good methods for indirect control of insects and diseases. These practices reduce the host material and breeding grounds of pests that may spread to healthy trees. In regions with a high incidence of a known pest, foresters attempt to avoid serious trouble by planting only trees known to resist existing pests in the regions where the trees are grown. Many forest genetic programs have as a major goal the selection and breeding of trees with insect and disease resistances.

Occasionally the natural conditions that suppress the population of pest organisms change, and outbreaks in forests may reach epidemic proportions. Even-aged stands and plantations with trees of the same species and of uniform size and age often create perfect conditions for the rapid spread of insects and diseases. Even the more complex uneven-aged forests with their inherent check-and-balance systems can develop devastating populations of pests. At this point the forester must consider direct control measures.

Because effective direct control of insects and diseases of standing timber is generally expensive, it is employed only when the potential mortality or loss in growth is extreme. Routine monitoring of insects and diseases allows foresters to schedule timely harvests of infested trees and to limit the spread of the problem to uninfested trees or areas. These sanitation and salvage harvests, coupled with piling and burning the limbs and branches left after logging, reduce the material and conditions that allow pest populations to develop. Traps baited with sex-attractant chemicals, or pheromones, are a promising method to reduce breeding populations of certain insects. Application of insecticidal or fungicidal sprays from the ground or from low-flying aircraft offer a short-term measure to check sudden plagues of insects or outbreaks of fungal diseases. Action has most frequently been taken against exceptional outbreaks of defoliating caterpillars, including those of the gypsy moth in the United States, the nun moth in central Europe, and the pine looper moth in England. At the time of year when feeding caterpillars are most vulnerable, light aircraft fly across the forest on carefully planned courses, distributing pesticides.

Control of defoliating caterpillars

By courtesy of Brandon Chaney

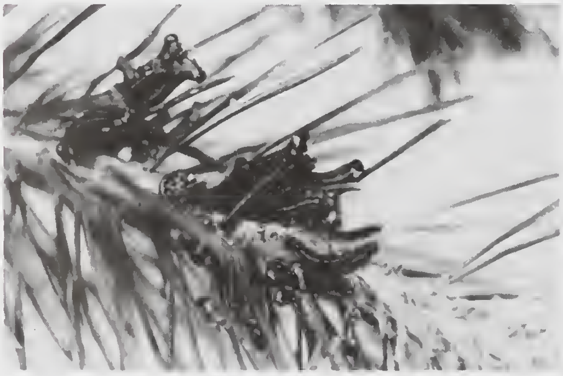


Figure 3: European conifer sawflies (*Neodiprion sertifer*) feeding on needles of Scotch pine.

A disadvantage of these blanket treatments by potent, broad-spectrum chemicals is that they also eliminate parasitic and predatory insects that serve as natural controls on the pest's numbers; they may also adversely affect birdlife. In practice, large-scale chemical treatments of forests are infrequent and are restricted to a small proportion of the areas at risk. Generally, natural control through predatory organisms, which also cycle opportunistically in a slightly delayed sequence with the pest populations, combined with physical factors like cold winters, provides adequate checks. Biological control involving the release of predators or diseases of pests is promising in some situations.

Less spectacular preventive measures are commonly taken as routine steps in practical forestry to lessen anticipated losses. Nursery stock, easily reached and handled, may be grown in fumigated seedbeds and sprayed during production so that uninfested and vigorous seedlings are planted in the forests. Prompt removal of logs from the forest to distant processing mills transfers beetles that may emerge from beneath the bark to areas where they can do no harm. Stumps of freshly felled conifers can be easily and cheaply treated by brushing on a fungicide to check the white root-rot fungus, *Fomes annosus*. This is a serious agent of decay that spreads underground through root grafts after gaining entry via the exposed surface of a felled stump.

Agroforestry. Agroforestry is a practice that has been utilized for many years, particularly in developing countries, and is now widely promoted as a land-use approach that yields both wood products and crops. Trees and crops may be grown together on the same tract of land in various patterns and cycles. The trees may be planted around the perimeter of a small farm to provide fuelwood and to serve as a windbreak. The limbs and foliage may be removed periodically for livestock fodder. Trees also may be planted in rows that alternate with crops or they may be planted more densely with interplanting of crops until crown closure of the trees precludes further crop production. These practices are most extensively used as a part of subsistence agriculture, but their use in large-scale production systems is becoming more common.

Urban forestry. Urban forestry, which is the management of publicly and privately owned trees in and adjacent to urban areas, has emerged as an important branch of forestry. Urban forests include many different environments such as city greenbelts; street and utility rights-of-way; forested watersheds of municipal reservoirs; and residential, commercial, and industrial property. An important distinction between urban and rural forestry is that urban trees are more highly valued than rural trees and often receive expensive individual care and attention. Many professional foresters are trained to handle the special problems of urban trees and to foster the diverse benefits they provide. (W.R.C./P.E.P./H.L.E.)

Wood production

Wood, botanically the principal strengthening and water-conducting tissue of stems and roots, is produced by many plants, including herbaceous ones, but wood valuable as a material, as considered in this article, derives mainly from the trunks of forest trees. As such, wood has been in service since humans appeared on Earth and has contributed to survival and to the development of civilization. In contemporary times, in spite of technological advancement and competition from metals, plastics, cement, and other materials, wood maintains a place in most of its traditional roles, and its serviceability is expanding through new uses with the result that its consumption is steadily increasing. The long list of present wood uses includes products in which its natural texture is retained and others in which the wood is mechanically and chemically modified to the extent that its presence cannot be recognized. In addition to well-known products, such as lumber, furniture, and plywood, wood is the raw material for wood-based panels, for pulp and paper, and many other products, especially chemical derivatives of cellulose and lignin. Finally, wood is still an important fuel in much of the world.

The versatility of wood is basically attributable to its structure, chemical composition, and properties. Produced by many botanical species, it is available in various colours and grain patterns. In relation to its weight, wood has high strength. It is insulating to heat and electricity and has desirable acoustical properties. Further, wood imparts a feeling of "warmth" not possessed by competing materials, such as metals, and is relatively easily worked. Cellulose is mostly obtained from wood. Wood is found throughout the world and is a renewable resource—in contrast to coal, ores, and petroleum, which are gradually exhausted. Wood has certain undesirable characteristics, however. It may burn and decay. It is hygroscopic (moisture absorbing), and in gaining or losing moisture it changes dimensions. As a biological product, wood is variable in quality.

STRUCTURE OF WOOD

Growth rings

General xylem structure and cell types. Examination of a stump or the cross section of a tree trunk reveals a series of successive growth layers of wood that surround a small central pith and are protected by a layer of bark. Between the bark and the wood is a narrow sheath of tissue called the vascular cambium. This lateral meristem is indistinguishable with the naked eye and produces all the cells that develop into xylem, or wood. In the temperate zones each growth ring of wood may be a product of one year's growth, but various environmental conditions may induce in each year the formation of more than one growth layer or discontinuous growth layers. In tropical areas growth rings are formed in response to wet and dry periods or other environmental cues. For these reasons, the term growth ring is preferred over annual ring.

Earlywood and latewood. Growth rings are visible because of distinct boundaries that result from the transition of earlywood to latewood. Earlywood cells are produced by the cambium early in the growth period and are typically less dense than the latewood cells because wider cells with thinner walls predominate in the earlywood (Figure 4). The latewood forms a distinct boundary between growth rings because of its sharp contrast to the earlywood of the following season. The transition from earlywood to latewood in the same growth ring is more or less gradual. The relative amounts of early- and latewood are affected by environmental conditions and differences in tree species. The proportion of early- and latewood affects the physical properties of wood and hence is an important concern related to utilization of wood derived from intensively managed tree plantations where growth conditions are altered.

Sapwood and heartwood. The sapwood is the portion of the xylem that stores food produced in photosynthesis and conducts water and dissolved nutrients to the crown of a tree. As a tree ages, the earlier produced growth rings of xylem usually become nonfunctional and develop into heartwood. Formation of heartwood involves the accumu-

From H.A. Core, W.A. Cote, and A.C. Day, *Wood Structure and Identification*, 2nd ed. (Syracuse: Syracuse University Press, 1979), by permission of the publisher

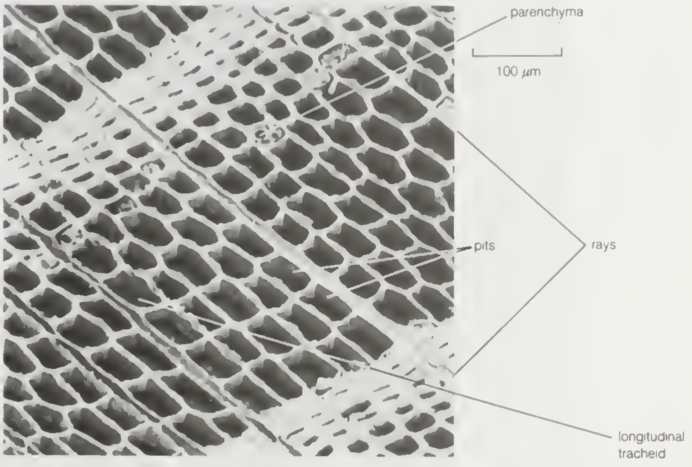


Figure 4: Scanning electron micrograph of a cross section of redwood (*Sequoia sempervirens*) showing earlywood-to-latewood transition in longitudinal tracheids. Rays, pits, and the parenchyma are also shown.

lation of metabolic by-products that may be inhibitory or even toxic to living cells. Movement of these substances occurs along the rays toward the centre of the tree where they eventually result in the death of living cells. Although sometimes pale, heartwood is generally darker in colour than the sapwood and is often preferred for many uses because the gums and resins deposited in the cell cavities and spaces between cells impart resistance to insects and decay, and they give the wood a rich colour.

Vascular rays

Transverse, radial, and tangential sections. Examination of a block of wood with the aid of low magnification reveals two distinct systems of cells. The axial system contains files of cells with their long axes oriented vertically in the stem, while the radial system comprises files of cells oriented horizontally with respect to the stem axis. Each of the two systems has its characteristic appearance in the three kinds of sections employed in the study of wood: transverse, radial, or tangential. The transverse, or cross section, cuts at right angles to the main axis of the tree to reveal the smallest dimension of cells in the axial system. The radiating pattern of vascular rays across growth rings is exposed. When a tree is cut lengthwise, either radial or tangential sections are obtained. Both show the vertical extent of cells of the axial system, but they give different views of the rays. The radial section passes through the pith and exposes the rays as horizontal bands lying across the axial system. Growth rings appear as parallel bands. A tangential section cuts at a tangent to the growth rings and cuts rays perpendicular to their horizontal extent to reveal the height and width. Growth ring arrangement is paraboloid in appearance. This occurs because the wood is cut at an angle due to the slight taper of the tree trunk. The different structural aspects of wood revealed in these respective sections accounts for the various grains or patterns seen in sawn lumber.

Cell types. Wood is a composite of tiny cells. Indicative of their small size is the estimate that one cubic metre of spruce wood contains 350,000,000,000–500,000,000,000 cells. Differences in cell form and arrangement account for an anatomy unique to each tree species. The principal cell types of wood are tracheids, vessel members, fibres, and parenchyma. Most cells are tubelike and are arranged parallel to the axis of the trunk. The tracheid is a primitive cell type with closed ends.

Evolutionary development of the xylem followed two lines that enhanced efficiency of conduction through development of vessel members and structural support through development of fibres. Both of these types of cells evolved from tracheids. Vessel members present wide variation in length, from 0.3 to 1.3 millimetres. Diameters range, in general, from about 0.01 to 0.5 millimetres. Pits in the sidewalls of vessel members generally are smaller than those in tracheids, but their function in translocation of water in the xylem is limited because of the large perforations in the end walls. Obstructions in vessel members called tyloses may occur and are frequent in heartwood. Fibres, too, are shorter than tracheids (one to two millimetres on the average) and have narrow diameters, closed ends, and thick walls. Cells with characteristics of more than one type of the basic cells occur in wood and are known as vascular tracheids, fibre tracheids, and libriform fibres.

Parenchyma cells are the simplest cell type in wood. They are blocklike and very small (0.1–0.2 millimetres in length). The epithelial cells that line resin canals and gum ducts are specialized parenchyma cells. Almost all wood cells in living trees are dead. Their protoplasmic contents have disintegrated, but the rigid cell walls remain around the cell cavities. The exceptions are a few rows of young cells next to the cambium that are produced during current growth and have not completed development and parenchyma cells located in the sapwood.

Structural variations and defects. The cellular composition and arrangement of wood varies among species. This influences appearance and properties and makes for a wide choice of woods for various uses. It also is the basis for wood identification. Woods of gymnosperm species, such as pine and spruce, are known as softwoods, whereas those of broad-leaved angiosperm species, such as oak and beech, are hardwoods. However, the implied distinction is not true; some hardwoods (*e.g.*, balsa) are softer than some softwoods (*e.g.*, yew).

Porous and nonporous wood

In terms of wood anatomy, trees are placed in two general categories—porous or nonporous. With only rare exceptions these anatomical categories correspond to the taxonomic classification of trees. Gymnosperms are nonporous and angiosperms are porous. Gymnosperm wood is relatively simple in structure. Its most striking feature is the absence of vessels and hence the name nonporous. The wood consists primarily of tracheids and fibres in the axial system. Axial parenchyma and resin canals are present in certain species, but radial parenchyma is always present and constitutes the rays, sometimes together with

From H A Core, W A Cote, and A C Day, *Wood Structure and Identification*, 2nd ed (Syracuse. Syracuse University Press, 1979), by permission of the publisher

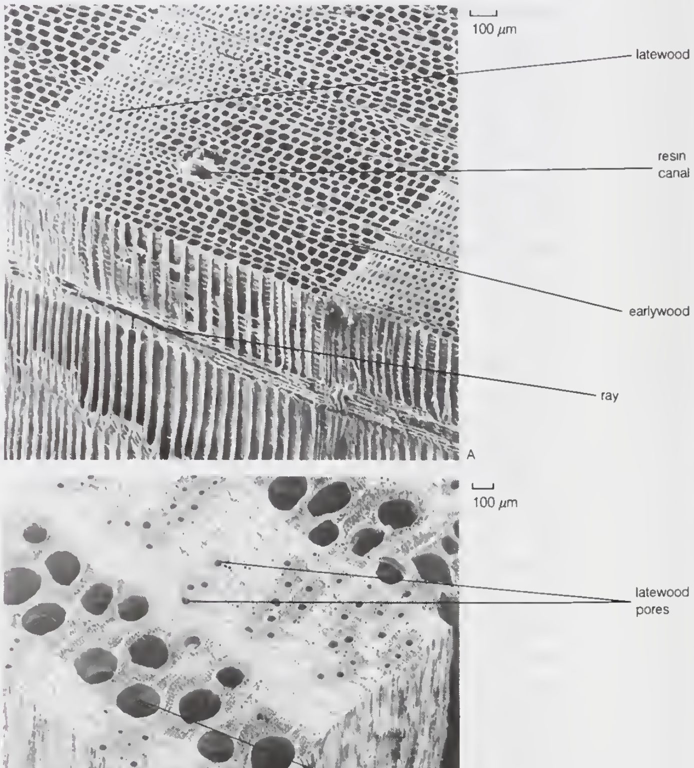


Figure 5: *Types of wood based on xylem structure as seen in scanning electron micrographs.*
(A) Nonporous wood of red pine (*Pinus resinosa*). A ray and a resin canal are also shown. (B) Ring-porous wood of red oak (*Quercus rubra*). (C) Diffuse-porous wood of aspen (*Populus grandidentata*).

radial tracheids (Figure 5A). The wood of angiosperm trees consists of vessels, tracheids, fibres, and parenchyma in various proportions. All four cell types occur in the axial system. The size and distribution of vessels allows for further classification as ring-porous and diffuse-porous woods. Ring-porous trees such as oak and ash have distinct vessels of larger diameter in the earlywood, whereas diffuse-porous trees such as birch and maple have vessels of about the same size uniformly distributed throughout a growth ring (Figure 5B,C). Parenchyma cells and occasionally tracheids constitute the radial system.

Variation in wood is caused by the presence of defects such as knots, spiral grain, compression and tension wood, shakes, and pitch pockets. Knots are caused by inclusion of dead or living branches as a tree grows in circumference. Spiral grain is the spiral arrangement of cell elements with regard to tree axis. Compression and tension wood are structural abnormalities in gymnosperms and angiosperms, respectively, that form when trees deviate from the normal, vertical position because of wind or other loads. Shakes are separations of wood tissue, and pitch pockets are separations filled with resin. Defects, according to kind and extent, may adversely affect the appearance, strength, dimensional stability, and other properties of wood.

Ultrastructure and chemical composition. Polarization microscopy, X rays, electron microscopy, and other techniques provide information regarding the structure of cell walls and other features hidden to light microscopes. Cell walls are crystalline. They are composed of a thin primary wall and a much thicker secondary wall, the latter made of three layers. The smallest visible building units of cell wall are the fibrils, which appear stringlike under the electron microscope. The orientation and weaving of microfibrils varies, which makes possible the distinction of layers. The secondary wall is absent in pit areas, and the primary wall, consisting of randomly and loosely arranged fibrils, forms the pit membrane. In tracheids of gymnosperms, the pit membrane possesses a central thickening, the torus.

The principal compound in cell walls is cellulose. Its molecules are linear chains of glucose, which may reach four microns in length. Orderly arrangement of cellulose molecules in fibrils (micelles) accounts for its crystalline properties. Noncellulosic constituents (hemicelluloses, lignin, and pectins) encrust the matrix among fibrils. Some hemicelluloses appear to serve as an important cross-link between the noncellulosic polymers and cellulose. Lignin is a complex substance that imparts rigidity to cell walls. Pectins are important constituents of the layer between cell walls (middle lamella).

Cellulose and the other chemical constituents are contained in wood in the following proportions (in percent of the oven-dry weight of wood): cellulose 40–45 percent (about the same in gymnosperms and dicotyledonous angiosperms); hemicelluloses 20 percent in gymnosperms and 15–35 percent in angiosperms; lignin 25–35 percent in gymnosperms and 17–25 percent in angiosperms; and pectic substances in very small proportion. In addition, wood contains extractives (gums, fats, resins, waxes, sugars, oils, starches, alkaloids, tannins) in various amounts (usually 1–10 percent, sometimes 30 percent or more). Extractives are not structural components but are deposited in cell cavities and intercellular spaces and may be removed (extracted) without change of wood structure.

Extractives

WOOD PROPERTIES

Density and specific gravity. Density is the weight or mass of a unit volume of wood, and specific gravity is the ratio of the density of wood to that of water. In the metric system, density and specific gravity are numerically identical; *e.g.*, the average density of the wood of Douglas fir is 0.45 grams per cubic centimetre (28 pounds per cubic foot) and its specific gravity 0.45, because one cubic centimetre of water weighs one gram. The density of wood varies from about 0.1 to 1.2 grams per cubic centimetre (specific gravity 0.1 to 1.2; Table 2). Differences between species and among samples of the same species are attributable to different proportions of wood substance, void volume (volume of cell lumens and wall spaces), and the content of extractives. The amount of extractives in wood varies from less than 3 percent to more than 30 percent of the oven-dry weight. It is obvious that the presence of these materials, located to a large extent within the cell wall, can have a major effect upon the density.

Determination of the density of wood in relation to that of other materials is difficult because wood is hygroscopic, and both its weight and volume are greatly influenced by moisture content. In order to obtain comparable figures, weight and volume are determined at specified moisture contents. The standards are oven-dry weight (practically zero moisture content) and either oven-dry or green volume (moisture content above fibre saturation point, which averages about 30 percent).

Green volume

Most mechanical properties of wood are closely correlated to density and specific gravity. It is possible to learn more about the nature of a wood sample by determining its specific gravity than by any other simple measurement.

Hygroscopicity. Wood is hygroscopic (*i.e.*, exhibits an affinity for water) and can absorb water as a liquid, if in contact with it, or in the form of vapour from the surrounding atmosphere. Though wood may absorb other liquids and gases, water is the most important. Because of its hygroscopicity, wood, either as a part of the living tree or as a material, always contains moisture. (Water and

Table 2: Properties of Certain Species of Wood

species	density*	percentage shrinkage				mechanical properties (kg/cm²)†							
		axial	radial	tangential	volume	static bending		compression		tension		hardness	
						modulus		parallel	perpendicular	parallel	perpendicular	parallel	perpendicular
						elasticity	rupture						
Lignum vitae (*Guaiacum officinale*)	1.23	0.1	5.6	9.3	15.0	123,000	—	1,260	900	—	—	1,970	—
Oak, white (*Quercus alba*)	0.71	0.9	5.3	9.0	15.2	125,000	1,070	520	93	—	56	690	620
Beech, red (*Fagus sylvatica*)	0.68	0.3	5.8	11.8	17.9	160,000	1,230	620	95	1,350	70	780	675
Birch, European (*Betula verrucosa*)	0.61	0.6	5.3	7.8	13.7	165,000	1,470	510	—	1,370	70	490	—
Pine, Scotch (*Pinus sylvestris*)	0.49	0.4	4.0	7.7	12.1	120,000	1,000	550	77	1,040	30	300	250
Douglas fir (*Pseudotsuga menziesii*)	0.45	0.3	4.2	7.4	11.9	115,000	790	470	65	1,050	24	320	270
Spruce, European (*Picea abies*)	0.43	0.3	3.6	7 8	11.9	110,000	780	500	58	900	27	270	160
Redwood (*Sequoia sempervirens*)	0.36	0.3	2.4	5.0	7.7	79,000	580	370	45	770	20	320	180
Balsa (*Ochroma lagopus*)	0.13	0.6	3.0	3.5	7.1	26,000	190	94	13	—	10	—	80

*Grams per cubic centimetre of oven-dried wood. †Based on small, clear specimens (moisture content 12%).
Source: Adapted from Kollmann, *Technologie des Holzes und der Holzwerkstoffe* (1951).

moisture are used here without distinction.) This moisture affects all wood properties, but it should be noted that only moisture contained in cell walls is important; moisture in the cavities merely adds weight.

The amount of moisture held in cell walls varies from about 20 to 35 percent (on the basis of oven-dry weight of wood). The theoretical point at which cell walls are completely saturated and cell cavities empty is known as the fibre saturation point. Beyond this point, moisture goes into the cavities, and when these are completely filled, the maximum moisture content of the wood is reached. Moisture content of some woods can be high. Very light woods, such as balsa, can hold up to about 800 percent, pine 250 percent, beech 120 percent, and so on.

When green wood is exposed to the atmosphere, its moisture content gradually decreases. Moisture in the cell cavities is lost first. In time the moisture content of wood falls to levels ranging (for localities in the temperate zones) from about 6 to 25 percent (average 12–15 percent). Local conditions of air temperature and relative humidity dictate the final moisture level.

Hygroscopicity is of primary importance because moisture in wood affects all wood properties. It has a direct relation to weight of logs and green lumber, with consequent influence on transportation costs. Dimensions change, as explained below (see *Shrinkage and swelling*). Resistance to decay and insects is greatly affected. Also influenced is processing, such as drying, preservative treatment, and pulping. Gluing and finishing and the mechanical, thermal, and acoustical properties of wood are all affected by its moisture content.

Shrinkage and swelling. Wood is subject to dimensional changes when its moisture fluctuates below the fibre saturation point. Shrinkage of the cell wall, and therefore of the entire wood, occurs as moisture escapes from between long-chain cellulose and hemicellulose molecules. These molecules can then move closer together. The amount of shrinkage that occurs is generally proportional to the amount of water removed from the cell wall. Swelling is simply the reverse of this process. It is characteristic that dimensional changes are anisotropic; *i.e.*, different in axial, radial, and tangential directions. Average values for shrinkage are roughly 0.2 percent, 4 percent, and 8 percent, respectively. Volumetric shrinkage averages 12 percent (Table 2). These values refer to changes from green to oven-dry conditions and are expressed in percent of green dimensions. The negligible longitudinal shrinkage of normal wood is one of the characteristics that makes lumber and lumber products such usable building materials. The factors, in addition to moisture content, that affect shrinkage and swelling are density (specific gravity), extractives, mechanical stresses, and abnormalities in wood structure.

Dimensional changes in wood caused by shrinkage and swelling may result in change of shape, checking (formation of cracks), warping, case hardening (release of stresses in resawing or other machining that causes warping), honeycombing, and collapse. Thus, the fact that wood shrinks and swells constitutes a great obstacle to its utilization. Several methods are used to improve the dimensional stability of wood, including resin impregnation to replace water in cell walls and treatment with various chemicals to eliminate the binding sites for water molecules. Though suitable for some commercial use, these are sufficiently expensive to limit their application to specialty items. Large-scale dimensional stabilization is accomplished by special wood construction. Plywood, discussed below, is an example.

Dimensional stabilization techniques

Mechanical properties. The mechanical or strength properties of wood measure its ability to resist applied forces that might tend to change its shape and size. Resistance to such forces depends on the magnitude and manner of application of the force. It also depends on various characteristics of the wood, such as moisture content and density. The term strength is often used in a general sense to refer to all mechanical properties. This can lead to confusion since there are many different types of strength and elastic properties. A wood that is relatively strong with respect to one strength property may rank lower in a different property when compared to another species. It is also important to note that wood has drastically different strength properties parallel to the grain than it does across the grain—*i.e.*, is anisotropic.

The mechanical properties of wood include strength in tension and compression (axial and transverse), shear, cleavage, hardness, static bending, and shock (impact bending, toughness). Respective tests determine stresses per unit of loaded area (at elastic limit and maximum load) and other criteria of strength, such as modulus of elasticity (a criterion of stiffness), modulus of rupture (bending strength), elastic resilience, and toughness. Tests are normally conducted with small, clear specimens, usually two by two inches or two by two centimetres in cross section. Laboratory data are analyzed to produce working stresses, which are available for use by engineers and architects in designing wooden structures. Mechanical properties of woods of several tree species are given in Table 2. Tests are sometimes conducted with structural components of actual size. Individual cells (tracheids, fibres) are also subject to testing since their strength is related to the strength of products (*e.g.*, paper).

Density is the best index of the strength of clear wood; higher density indicates greater strength. The strength of wood is also influenced by its moisture content when it fluctuates below the fibre saturation point. Generally, a decrease in moisture content is accompanied by an increase in most strength properties. Temperature and duration of loading also affect strength. In general, strength falls as temperature rises. Wood loaded permanently will support smaller loads than those indicated by a short-time test in the laboratory. The most important strength-reducing factors are wood defects, such as knots, abnormal anatomy resulting from compression and tension, and grain deviations. Their adverse effect depends on the kind and extent of the defect, position, and manner of loading.

Lumber grading

Defects constitute the basis for visual grading rules of lumber and other wood products. The rules for grading hardwood lumber are different than those for softwoods. Hardwood lumber grades were developed with the assumption that the lumber would be cut into smaller pieces for the manufacture of furniture or millwork parts. Grading is based upon the percentage of the board that is usable in smaller clear pieces free of defect on one or both sides. The regular hardwood lumber grades are thus not readily adaptable to applications where the entire board will be used as a single piece.

Softwood lumber grades, in contrast, are based principally on structural uses for which the strength or appearance of the entire board is important. Grades are usually assigned by a lumber grader, who must make a rapid judgment about the strength-reducing defects present in each piece. Most structural lumber is graded in this way. Grading leads to more efficient utilization of wood and is essential in order to achieve adequate standards of safety in wooden structures.

Thermal properties. Although wood expands and contracts with varying temperature, these dimensional changes are small in comparison to the shrinkage and swelling caused by variation of moisture content. In most cases expansion and contraction are negligible and without practical importance; only temperatures below 0° C (32° F) may cause surface checks, and in living trees, unequal contraction of outer and inner layers may result in frost cracks. Such low thermal expansion and contraction, in conjunction with low heat conductivity, constitute advantages of wood when it is exposed to fire. In addition, the low heat conductivity of wood (high insulating value) makes it desirable for building construction. Heat conductivity is about two to two and a half times greater axially than transversely and increases with density and moisture content.

Exposed to high temperatures, wood burns; at a temperature of about 400° C (752° F) wood ignites easily because of the production of flammable gases. The heating value of one kilogram of dry wood is about 4,000–5,000 kilocalories. Differences among species derive from differences in density and the presence of extractives (*e.g.*, resin in pines).

Electrical properties. Very dry, especially oven-dry, wood constitutes an excellent insulator. As moisture content increases, however, electric conductivity increases; and the behaviour of saturated wood approaches that of water. Noteworthy is the spectacular decrease of electric resistance when moisture content increases from zero to fibre saturation point. Within this range, electric resistance decreases about 10,000,000 times, whereas from fibre saturation point to maximum moisture content it decreases only about 50 times or less. Other factors, such as species and density, have little effect on the electric resistance of wood.

Dielectric property of wood

Important also is the dielectric property of wood. Wood is a nonconductor of electricity but will sustain the force of an electric field passing through it. This property, expressed in dielectric constant and power factor, assumes a practical importance with wood in drying, gluing (with high-frequency electric current), or making electrical meters (capacity and radio-frequency power-loss type) for measuring its moisture content.

Wood exhibits the piezoelectric effect; *i.e.*, electric polarization occurs under mechanical stress and also mechanical strain in an electric field.

Acoustic properties. Wood can produce sound (by direct striking) and can amplify or absorb sound waves originated from other bodies. For these reasons, it is a unique material for musical instruments and other acoustic purposes. The pitch of sound produced depends on the frequency of vibration and the dimensions, density, moisture content, and modulus of elasticity of the wood. Larger dimensions, lower moisture content, and higher density and elasticity produce sounds of higher pitch (more acute in tone). Sound waves originating from other bodies and striking wood are partly absorbed and partly reflected; wood is also set in vibration. The sound may be amplified, as in violins, guitars, organ pipes, and other musical instruments, or absorbed, as in wooden partitions. Normally, wood absorbs only a small portion of acoustic energy (3–5 percent), but special constructions with empty spaces and porous insulation boards may increase this capacity up to 90 percent. The velocity of sound in wood is high (3,500–5,000 metres per second axially and 1,000–1,500 metres per second transversely). Defects such as decay affect acoustic properties; use of this fact is made in nondestructive testing of wood.

HARVESTING THE WOOD CROP

A prerequisite to harvest is a management plan, which determines the yearly yield and the method of removal. The harvest method chosen can involve clear-cutting large areas or selective cutting of individual trees or groups of trees. For a forest harvested under the sustained-yield concept the volume of timber removed at periodic time intervals is dependent on the net growth of all trees during that interval. This concept, combined with natural and artificial seeding and planting, ensures a continuous production of wood and conservation of forests. The season of harvest is not determined by the time of ripening, as with agricultural crops, but by such factors as the conditions of work for personnel, machines and animals, and the danger of damage to the remaining forest and to the harvested wood. Because felled trees are vulnerable to attack by fungi and insects, the harvest is timed to avoid conditions favourable for these organisms.

Harvesting includes felling, bucking (cross-cutting into logs), limbing, debarking, and skidding (*i.e.*, moving the logs from the felling location) to the roadside or concentration yard, from where the logs are transported to industries. Felling is commonly accomplished by chain saw; the ax and handsaw are little used today. The chain saw is also used for bucking and in most cases for limbing. Bucking is not always done at the felling site. Sometimes whole trees are skidded to a concentration yard for further processing. Debarking is sometimes done in the forest by ax or spud (a combination spade and chisel) or by portable mechanical debarkers, and sometimes in factories by stationary mechanical debarkers or water jets. Special equipment has been developed for harvesting pulpwood (Figure 8). An example is a combine harvester equipped with giant scissors, which shear the tree at the base. It is then lifted to a carriage and drawn through a trimmer, which strips off the branches. Another blade bucks the trunk into logs, which fall into an attached cradle. Sometimes the entire tree, including its branches, is chipped in the woods. The chips are blown directly into a truck or are carried by pipeline to a pulp mill.

Pulpwood harvesting

Skidding is done by tractors or by animals; in various forests of the world, horses, mules, oxen, and elephants are employed. Tractors are usually employed in combination with steel cables, and the logs are skidded on the ground or lifted partially or wholly off the ground. In the northwest United States tall trees, 80–100 metres high, are topped by a climbing logger and are employed as masts, or spar trees, to attach cables for skidding. In rare cases where slopes are steep and erosive, helicopters or giant balloons are used. In general, mechanization of harvesting operations is the trend, but regions of small annual yield and unfavourable topography restrict the potential of expensive machines, and in many countries human and animal labour is still commonly used.

The main source of usable wood is the tree trunk. Tree tops and heavy branches are cut into short lengths for cordwood and stacked wood, or they are chipped. Stumps, which should be as low as possible, roots, logging residues, and bark, if logs are debarked, remain in the forest. Machines also have been designed to extract the roots of the trees. This technique is employed to some degree in the coniferous forests of the southeastern United States.

WOOD UTILIZATION

This section is concerned with the main products of primary processing of wood, and related treatments, such as drying and preservation, that ensure its better performance in use. Some of these products, such as poles, posts, and railroad ties, are used directly, but most constitute intermediate materials that by further processing are manufactured into final products or structures.

Roundwood products. Poles, posts, and certain mine timbers are products in round form. Poles are used in telecommunication lines (telegraph, telephone) or as pilings (*i.e.*, foundations for wharves or buildings), and posts in fences, highway guards, and various supports. As a rule, these products are subjected to preservative treatment. The bark is removed in the forest or factory.

Sawn wood. Lumber is the main sawn-wood product. Lumber of large dimensions (more than about 10 centimetres in width and thickness) suitable for heavy constructions is called timber. This term, however, is also loosely applied to wood of a forest stand and to products of round form. Another important product made by sawing, and sometimes by hewing, is railroad ties.

Processing logs into lumber

Lumber is usually the product of the sawmill and is produced in varying sizes (usual, approximate dimensions: thickness two to 10 centimetres, width eight centimetres and over, length two to six metres). Conversion of logs to lumber involves breakdown into boards of various thicknesses, resawing, ripping, and crosscutting. The organization of production varies by manufacturing plants, but a generalized scheme is as follows. Logs, transported from the forest, are stored in water, usually a pond or river, or in a ground storage yard. If a long ground-storage time is anticipated, the logs are kept under a constant water mist. Each log enters the mill on a conveyor; in large operations it is mechanically debarked and in some is crosscut to length. Supported on a carriage, it is brought to a headsaw, which may be of three types: band saw, gang saw, or circular saw. A band saw consists of an endless band of steel, equipped with teeth usually on one edge only and moving around two wheels—one powered and the other free-running. Gang saws commonly consist of a reciprocating (vertical or horizontal) frame in which a number of saw blades are mounted at predetermined lateral distances. A circular saw consists of a circular blade having teeth (sometimes removable) on its periphery and mounted on a shaft. Band and gang saws have relatively thin blades and are therefore less wasteful than circular saws; band and circular saws permit changing board thickness and turning of the log after each cut; therefore, breakdown is more advantageous in terms of yield and grade. In general, and except for logs of very large diameter, gang saws are extensively used, especially in Europe, as headsaws for softwoods, and band saws are used for hardwoods.

Breakdown is accomplished in one or more operations by the use of one, two, or more machines in combination; *e.g.*, if two gang saws are used, the first saw removes slabs (the outside pieces cut from a log) and, in certain cases, some boards. The piece produced is then turned 90° and introduced into the second saw, which converts it to boards. The second operation may be considered resawing; in general, resawing consists of either dividing thick boards into thinner ones or producing boards from slabs. Ripping, or edging, is the removal of wane or bark from the sides of boards by passing them through a band saw or a machine that has two small circular saw blades mounted on a shaft; one is stationary and the other may move sidewise, thus setting board width. Finally, certain boards are crosscut to square their ends and remove defects. Modern sawmills are controlled with the aid of computers and other types of electronic equipment.

Drying. Drying is an essential preparation of lumber and wood. Proper drying reduces the magnitude of dimensional changes due to shrinkage and swelling, protects wood from microorganisms, reduces weight and transportation costs, prepares wood better for most finishing and preservation methods, and increases its strength. Drying is accomplished in the open air or in kilns. Other special methods of drying also exist.

The air-drying yard

The air-drying yard is located close to the lumber plant on a dry site, where air movement is not obstructed by tall trees or buildings. The ground surface is kept free from debris and vegetation. The time required to air dry from green to 20 percent moisture content varies in general from about 20 to 300 days for material 2.5 centimetres thick, depending on species, place, and time of the year.

Air drying can be accelerated by means of fans, the application of "solar heat," and "predriers" (fans and low-temperature heating). With beech, walnut, and some other woods, steaming is also employed before air drying; this reduces drying time by increasing the rate of drying and at the same time darkens the wood and makes it more desirable for use as furniture.

Kiln-drying is conducted in a closed chamber, under artificially induced and controlled conditions of temperature, relative humidity, and air circulation. This method permits much faster reduction of moisture content to levels independent of weather conditions: reduction of moisture from 20 to 6 percent is accomplished in two to 15 days, and from green to 6 percent in two to 50 days for material 2.5 centimetres thick. The source of heat is usually steam circulating in pipe coils. Relative humidity is controlled by allowing steam to enter the chamber through a perforated pipe; such control is necessary in order to regulate the exit of moisture and avoid defects such as checking (cracking or splitting). In order to obtain satisfactory results air movement is necessary to carry the heat from its source to the lumber and to carry away the evaporated moisture; air circulation is produced by means of fans located within the kiln, and sometimes by blowers placed outside. Kiln-drying normally involves temperatures below 100° C (212° F), usually in the range 40°–75° C (104°–167° F). Such temperatures are high enough to kill insects, which is another advantage of kiln-drying over air drying.

In addition, wood may be dried by special methods that include the use of solvents, vapours, and chemicals and high-temperature drying, boiling in oily liquids, vacuum drying, electric drying, and drying by infrared radiation. The last three methods are expensive and therefore not commercially applicable.

Fungi

Preservation. Wood is subject to deterioration by fungi (causing stain and decay), insects, marine organisms, fire, and other destructive agents. By far the most important cause of wood loss is decay. Wood decays if the conditions are suitable for the growth and activity of fungi. Such conditions include favourable moisture, air, and temperature. A moisture content below 20 percent inhibits growth of fungi. If wood is kept under water, it also cannot be attacked because of the lack of sufficient available oxygen. There are many examples of wooden structures lasting hundreds or even thousands of years; *e.g.*, in the tombs of pharaohs. Under conditions favourable for fungi (for example, when wood is used in contact with soil) no wood is immune, although some species do better than others. The durability of exposed timber, such as railroad ties or mine timbers, can be greatly increased by impregnation with toxic chemicals. The application of preservatives is accomplished by brushing, spraying, dipping, steeping, hot and cold bath, and diffusion, but impregnation under pressure is the most efficient. This requires special installations such as treating cylinders (up to 50 metres in length and three metres in diameter, storage tanks, pumps, a boiler plant (for steam production), and other auxiliary equipment.

Veneer. Veneer is a thin layer, or sheet, of wood that is uniform in thickness—commonly about 0.6–eight millimetres. According to the method of production (see Figure 6), it is classified into rotary cut (cut in a lathe by rotating a log against a knife, similar to peeling), sliced (cut sheet by sheet from a log section or "flitch"), and sawn (produced by sawing with a special tapered saw). More than 90 percent of all veneer is rotary cut, but figured woods producing veneer for furniture and other decorative purposes are sliced; sawn veneer is seldom produced because it is a wasteful operation.

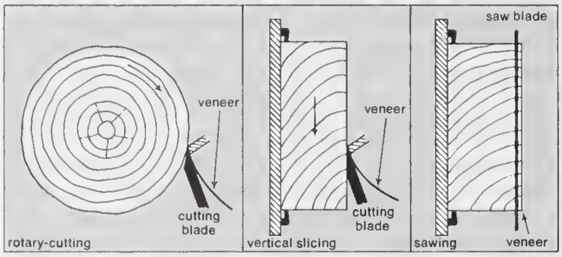


Figure 6: Methods of veneer production.

Veneers are used primarily for plywood and furniture, but they are also used in toys, containers of various kinds, matches, battery separations, and other products. An experimental product related to veneer is slice wood; this is thicker, and its production is less wasteful than lumber of the same thickness.

Plywood and laminated constructions. Plywood and laminated constructions are glued-wood products. Although gluing is an old art, practiced since ancient times, the modern development of various products was made possible by the improvement of glues—especially by the production of synthetic resin adhesives.

Manufacturing plywood

Plywood is a panel product manufactured by gluing together one or more veneers to both sides of a veneer, solid wood, or reconstituted wood core (see Figure 7). In the case of solid-wood-core plywood and reconstituted-wood-core plywood, an additional intermediate step is the production of cores, which are made by lateral gluing of blocks or strips of wood or by gluing oriented wood chips or flakes with resin adhesives.

In plywood the grain of alternate layers is crossed, in general at right angles; species, thickness, and grain direction of each layer are matched with those of their opposite number on the other side of the core; usually the total number of layers is odd (three, five, or more). Thus assembled, the panels are brought to presses for gluing with either natural (animal, casein, soybean, starch) or synthetic resins, such as phenol- or urea-formaldehyde. Certain synthetic resins, such as phenol-formaldehyde, properly used, may produce joints more durable than the natural wood itself—highly resistant to weather, microorganisms, cold, hot and boiling water, and steam and dry heat; such plywood is known as exterior plywood (in contrast to interior).

Plywood has many advantages over natural wood, an important one being greater dimensional stability. Its uniformity of strength, resistance to splitting, panel form, and decorative value make it adaptable to various uses. In addition to flat panels, plywood is manufactured in curved form (molded plywood), used for boats, furniture, and many other products; this is done by bending and gluing veneer sheets in one operation—by use of curved forms in

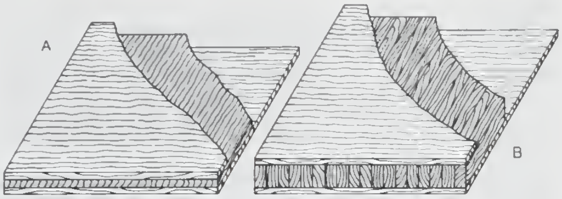


Figure 7: *Types of plywood.*
(A) Three-ply all-veneer. (B) Three-ply solid wood core (laminboard).

a press or by fluid pressure applied with a flexible "bag" or "blanket" of an impermeable material.

Another important glued product is laminated wood. This is produced mainly from lumber with the grain of all boards parallel to one another. The product is used in beams, columns, and arches for buildings, boat keels, aircraft carrier decking, minesweepers, and helicopter propellers. In curved products, production involves simultaneous bending and gluing. Laminated wood possesses several advantages over solid wood. Large members of various sizes and shapes impossible to make from solid wood can be fabricated; the individual boards used, due to their relatively small thickness, may be properly dried without checking, and defects, such as knots, may be removed; structures may be designed on the basis of required strength, and wood of low grade can be positioned accordingly; gluing permits utilization of small dimensions.

"Improved wood"

Special products are made of veneer sheets that are impregnated with synthetic resin, assembled parallel or in the conventional plywood manner, then pressed and glued. This results in "improved wood," characterized by high density; improved dimensional stability, strength, and appearance; and resistance to fungi, insects, fire, and weathering. Finally, veneer and plywood are fabricated into sandwich constructions by gluing in combination with other materials, such as fibreboard, paper, cloth, asbestos, metal, and plastics.

Particle board. This panel product is manufactured of particles of wood glued together. Particles are flakes, shavings, or splinters produced by cutting or breaking. Boards are produced either with the same particle geometry throughout or in layers—with different particle patterns on the faces and in the core.

Particle board production is a relatively new industrial development, begun in the early 1940s and rapidly expanding since in all countries. It was made possible by the development of synthetic resins and has greatly contributed to better wood utilization by permitting the use of residues of other wood-using industries and of harvesting operations in forests. Debarking is not always necessary.

Fibreboard. This panel product is made of fibres of wood. As in the case of particle board, unbarked wood of low quality may be utilized through pulp preparation, sheet formation, pressing, and finishing treatment. Pulping is usually mechanical or semichemical, but a so-called explosion process is also used, in which chips are subjected to high-temperature steam in a "gun" or high-pressure vessel; ejected through a quick-opening valve, they are reduced to pulp. Two basic processes—wet felting and air felting—are employed in sheet formation. Before entering into sheet formation, certain materials are added to the pulp to improve water resistance, strength, and other properties. In wet felting, glue (synthetic resin) is usually not used, though in the dry process it is added in a proportion of 1–4 percent (on the basis of dry fibre mass). Other additives, such as rosin, paraffin, wax, and chemicals, are used to increase resistance to microorganisms, insects, and fire. Air felting offers advantages with regard to water consumption and pollution and, for this reason, is preferred over the wet process. Air felting produces either semidry wood (12–45 percent moisture content) or dry wood (8–10 percent).

In general, panel products (plywood, particle, and fibreboard) serve a wide range of uses: building construction, including walls, floors, roofs, and doors; exterior siding, interior finishing (*e.g.*, wall paneling), and shelves; furniture; shipbuilding; automobile manufacture; refrigeration cars; toys; concrete formwork; and many others. Special types combine decorative value with thermal- and sound-conditioning properties.

Pulping processes

Pulp and paper. Wood is the main source of pulp and paper. Preliminary production steps are debarking and chipping. Pulping processes are of three principal types: mechanical, or grinding; chemical, or cooking with chemicals added; and semichemical, a combination of heat or chemical pretreatment with subsequent mechanical reduction. The yield of pulp ranges from about 40 percent by chemical methods to 95 percent by mechanical ones. Chemical processes are based on either acids (sulfite pulping) or alkalies (alkaline pulping, including the soda and the sulfate process). The pulp produced is washed, screened, thickened by the removal of most water, and bleached. Paper manufacture involves beating the pulp, loading, introducing various additives, refining, and running the pulp into the paper machine (see INDUSTRIES, CHEMICAL PROCESS: *Papermaking*).

Other products. The total number of products made of wood and its derivatives (*e.g.*, cellulose) is enormous, according to some estimates as high as 10,000.

Mechanically derived products. In addition to those already mentioned, some of the principal applications of wood include agricultural tools, aircraft, artificial limbs, barrels, baseball bats, baskets, blackboards, blinds (Venetian), bobbins, bowling pins, brush blocks, caskets, clothespins, crates, dowels, excelsior (wood-wool), fishing rods, golf clubs, gun stocks, handles, ice-cream spoons, insulating pins, ironing boards, ladders, novelties, oars, pallets, patterns and models, parquet flooring, paving blocks, pencils, picture frames, rules, scaffolding, scientific instruments, shingles, shoe lasts, skis, sleighs, smoking pipes, spools, tanks, tennis rackets, tongue depressors, toothpicks, vats, wood flour, and many others.

Chemically derived products. These include acetic acid, acetone, cellophane, cellulose acetate, charcoal, dyestuffs, ethyl alcohol, explosives, furfural, lacquer, methanol, molasses, oils, paper products, photographic films, pitch, plastics, rayon, rosin, sugars, synthetic sponges, tannins, tar, turpentine, vanillin, yeast, and many others. Promising developments include ablative carbon composite, a substance resistant to very high temperatures (2,800° C [5,072° F]); structural graphite composite—stronger, lighter, and more rigid than advanced metal alloys, suitable for airplane construction; and permanently fire-retarding rayon. The list of chemicals that can theoretically be produced from wood is a long one indeed and not materially different in length and contents from the list of products that can be made from petroleum.

The use of wood for fuel is diminishing in industrialized countries because of the increased use of petroleum; gas; electricity generated by solar, wind, atomic, and water

By courtesy of the Embassy of Finland

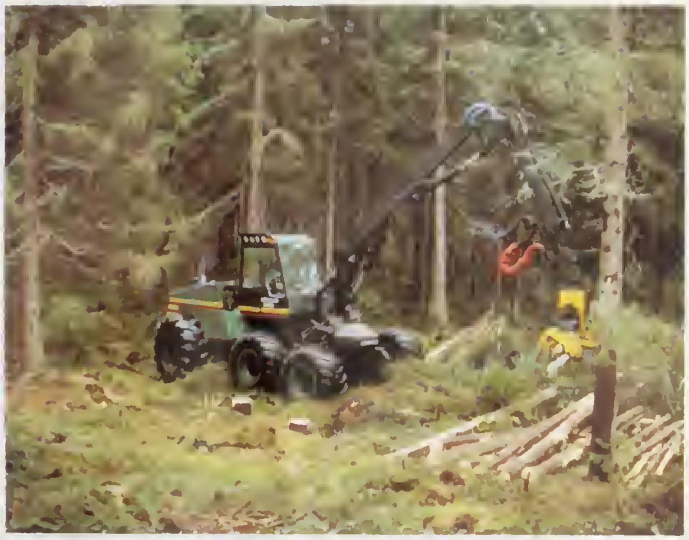


Figure 8: Single grip harvester at work in a tree plantation in Finland.

power; and coal. During periods of petroleum shortages wood is used for home heating, for generation of electricity via direct combustion, in the production of ethanol, and as a mixture with high-sulfur coal. Some countries are dependent on wood as the primary or sole source of fuel for cooking and heating. The production of fuelwood (from large coppice and low-quality forests) is essential to the survival of vast numbers of people, yet the problem of wood shortages is becoming more acute; this contributes to more complete utilization, which in turn depletes the existing wood resource and its potential for regeneration.

In addition to wood, increased attention is being focused on bark, which constitutes 10–15 percent of a tree volume. Bark is used in charcoal, as cork, in fibreboard (in mixture with wood), for soil improvement, and as a source of tannins and other chemicals. But further possibilities are recognized, particularly with respect to its chemical utilization. (G.Ts./W.R.C./P.E.P.)

BIBLIOGRAPHY

Forestry: Forestry textbooks usually deal with a specialized aspect of the science, often in one geographic region. WILLIAM M. HARLOW, ELLWOOD S. HARRAR, and FRED M. WHITE, *Textbook of Dendrology, Covering Important Forest Trees of the United States and Canada,* 6th ed. (1979), provides descriptions of major tree species in the region. Forest ecology is the subject of K.A. LONGMAN and J. JENÍK, *Tropical Forest and Its Environment,* 2nd ed. (1987); J.J. LANDSBERG, *Physiological Ecology of Forest Production* (1986); and HERMAN H. SHUGART, *A Theory of Forest Dynamics: The Ecological Implications of Forest Succession Models* (1984). For a general historical overview of forestry and related fields, see RICHARD C. DAVIS (ed.), *Encyclopedia of American Forest and Conservation History,* 2 vol. (1983).

Principles of forest management, worldwide in application, are systematically outlined by WILLIAM A. LEUSCHNER, *Introduction to Forest Resource Management* (1984); and JOSEPH BUONGIORNO and J. KEITH GILLESS, *Forest Management and Economics: A Primer in Quantitative Methods* (1987). KARL F. WENGER (ed.), *Forestry Handbook,* 2nd ed. (1984), is a reference book of data and methods in all aspects of forestry and allied fields. GRANT W. SHARPE, CLARE W. HENDEE, and WENONAH F. SHARPE, *Introduction to Forestry,* 5th ed. (1986); and CHARLES H. STODDARD and GLENN M. STODDARD, *Essentials of Forestry Practice,* 4th ed. (1987), give a complete overview of modern multiple-use forestry. The requisite elements of forest inventory are detailed in BERTRAM HUSCH, CHARLES I. MILLER, and THOMAS W. BEERS, *Forest Mensuration,* 3rd ed. (1982). Financial implications are studied in G. ROBINSON GREGORY, *Resource Economics for Foresters* (1987). The leading international sources of statistics on forestry and timber output are the publications of the Food and Agriculture Organization of the United Nations: *World Forest Inventory* (irregular), *FAO Forestry and Forest Product Studies* (irregular); and the special reports *Forest Resources of Tropical Africa,* 2 vol. (1981), *Forest Resources of Tropical Asia* (1981), and *Forestry in China* (1982). Fundamental studies of the physical bases of forest distribution and yield are given in *World Resources 1986* (1986), a report prepared by the World Resources Institute and the International Institute for Environment and Development.

The theory and practice of raising and tending tree crops are treated in such manuals as THEODORE W. DANIEL, JOHN A. HELMS, and FREDERICK S. BAKER, *Principles of Silviculture,* 2nd ed. (1979); and DAVID M. SMITH, *The Practice of Silviculture,* 8th ed. (1986), discussing temperate-zone forests. Details of handling seed and young stock are treated in R.L. WILLAN (comp.), *A Guide to Forest Seed Handling: With Special Reference to the Tropics* (1985); U.S. DEPARTMENT OF AGRICULTURE, *Woody-Plant Seed Manual* (1948); and MARY L. DURYEA and THOMAS D. LANDIS (eds.), *Forest Nursery Manual: Production of Bareroot Seedlings* (1984). Genetic improvement is discussed in KLAUS STERN and LAURENCE ROCHE, *Genetics of Forest Ecosystems* (1974); and M.N. CHRISTIANSEN and CHARLES F. LEWIS (eds.), *Breeding Plants for Less Favorable Environments* (1982). The intensive culture of forest plantations is discussed in BRUCE J. ZOBEL, GERRIT VAN WYK, and PER STAHL, *Growing Exotic Forests* (1987); and W.E. HILLIS and A.G. BROWN (eds.), *Eucalypts for Wood Production* (1978, reprinted 1984). Management of forest soils, a primary concern in intensive silviculture, is described in WILLIAM L. PRITCHETT and RICHARD F. FISHER, *Properties and Management of Forest Soils,* 2nd ed. (1987); and PEDRO A. SANCHEZ, *Properties and Management of Soils in the Tropics* (1976). Manipulation of soil fertility and study of soil microorganisms are presented in G.D. BOWEN and E.K.S. NAMBIAR (eds.), *Nutrition of Plantation Forests* (1984); ROBERT L. TATE, III, and DONALD A. KLEIN (eds.), *Soil Reclamation Process: Microbiological Analyses and Applications* (1985); and J.C. GORDON and C.T. WHEELER (eds.), *Biological Nitrogen Fixation in Forest Ecosystems: Foundations and Applications* (1983). The care of tree crops in tropical jungles of both hemispheres is outlined in I.T. HAIG, M.A. HUBERMAN, and U. AUNG DIN, *Tropical Silviculture* (1958). A wide range of Asiatic conditions is discussed in HARRY G. CHAMPION and S.K. SETH, *General Silviculture for India* (1968).

The management of forests as watersheds and the impact of forestry activities on water quantity and quality are discussed in WILLIAM E. SOPPER and HOWARD W. LULL (eds.), *Forest Hydrology: Proceedings of a National Science Foundation Advanced Science Seminar* (1967); H.C. PEREIRA, *Land Use and Water Resources in Temperate and Tropical Climates* (1973); and K.W.G. VALENTINE, *Soil Resource Surveys for Forestry* (1986). Nutrient losses from disturbed watersheds and the potential for accelerated loss attributable to acid deposition are examined in F.E. CLARK and T. ROSSWALL (eds.), *Terrestrial Nitrogen Cycles: Processes, Ecosystem Strategies, and Management Impacts* (1981); and S. BEILKE and A.J. ELSHOUT (eds.), *Acid Deposition* (1983).

Forest protection is treated in textbooks discussing specific hazards. ARTHUR A. BROWN and KENNETH P. DAVIS, *Forest Fire: Control and Use,* 2nd ed. (1973), outlines fire dangers and methods of control. T.T. KOZLOWSKI and C.E. AHLGREN (eds.), *Fire and Ecosystems* (1974); and HENRY A. WRIGHT and ARTHUR W. BAILEY, *Fire Ecology, United States and Southern Canada* (1982), discuss the environmental interactions associated with fire. JOHN S. BOYCE, *Forest Pathology,* 3rd ed. (1961); and ROBERT O. BLANCHARD and TERRY A. TATTAR, *Field and Laboratory Guide to Tree Pathology* (1981), give details of fungal diseases, climatic dangers, and airborne fume and salt damage. See also CARL F. JORDAN (ed.), *Amazonian Rain Forests: Ecosystem Disturbance and Recovery* (1987). Insect pests are described in ALAN A. BERRYMAN, *Forest Insects: Principles and Practice of Population Management* (1986).

Wood production: The structure and properties of wood are treated in F.W. JANE, *The Structure of Wood,* 2nd ed., rev. by K. WILSON AND D.J.B. WHITE (1970); GEORGE TSOUMIS, *Wood as Raw Material: Source, Structure, Chemical Composition, Growth, Degradation, and Identification* (1968), mainly on structure, with keys for identification of North American and European commercial woods; A.J. PANSHIN and CARL DE ZEEUW, *Textbook of Wood Technology: Structure, Identification, Properties, and Uses of the Commercial Woods of the United States and Canada,* 4th ed. (1980); FRANZ P. KOLLMANN and WILFRED A. CÔTÉ, JR., *Principles of Wood Science and Technology,* vol. 1, *Solid Wood* (1968), on structure, chemical composition, and biologic deterioration, with detailed coverage of properties; ALFRED J. STAMM, *Wood and Cellulose Science* (1964), fundamental information on properties; H.E. DESCH, *Timber: Its Structure and Properties,* 6th ed., rev. by J.M. DINWOODIE (1981), including a discussion of drying, preservation, and grading; and JOZSEF BODIG and BENJAMIN A. JAYNE, *Mechanics of Wood and Wood Composites* (1982). JOHN G. HAYGREEN and JIM L. BOWYER, *Forest Products and Wood Science* (1982), provides basic information concerning physical and chemical properties of wood and the nature of major wood products. H.A. CORE, W.A. CÔTÉ, and A.C. DAY, *Wood Structure and Identification,* 2nd ed. (1979), contains photomicrographs of wood structure.

Basic principles of tree felling and timber haulage are covered in STEVE CONWAY, *Logging Practices: Principles of Timber Harvesting Systems,* rev. ed. (1982); GEORGE STENZEL, THOMAS A. WALBRIDGE, JR., and J. KENNETH PEARCE, *Logging and Pulpwood Production,* 2nd ed. (1985); and ROELOF A.A. OLDEMAN (ed.), *Tropical Hardwood Utilization: Practice and Prospects* (1982). Methods and machines evolve so rapidly that recent issues of a trade periodical, such as *World Wood* (bimonthly), should be consulted for current practice and equipment. An interesting historical overview of the industry is provided in KENNETH L. SMITH, *Sawmill: The Story of Cutting the Last Great Virgin Forest East of the Rockies* (1986).

Works on the commercial utilization of wood include A.J. PANSHIN *et al., Forest Products: Their Sources, Production, and Utilization,* 2nd ed. (1962), which provides general information on all products; FRANZ P. KOLLMANN and WILFRED A. CÔTÉ, JR., *Principles of Wood Science and Technology,* vol. 2, *Wood Based Materials* (1975); DARREL D. NICHOLAS (ed.), *Wood Deterioration and Its Prevention by Preservative Treatments,* 2 vol. (1973); FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS, *Plywood and Other Wood-Based Panels* (1966); and KENNETH W. BRITT (ed.), *Handbook of Pulp and Paper Technology,* 2nd rev. ed. (1970). Certain products (and properties) are discussed in the U.S. DEPARTMENT OF AGRICULTURE, *Wood Handbook: Wood as an Engineering Material,* rev. ed. (1987); and GERMAN GURFINKEL, *Wood Engineering,* 2nd ed. (1981), containing data for use in design and specification.

(W.R.C./P.E.P./H.L.E./G.Ts.)

France

Historically and culturally among the most important nations in the Western world, the French Republic (République Française) has also played a highly significant role in international affairs, with former colonies in every corner of the globe. Bounded by the Atlantic Ocean and the Mediterranean Sea, the Alps and the Pyrenees, France has long provided a geographic, economic, and linguistic bridge joining northern and southern Europe. It is Europe's most important agricultural producer and one of the world's leading industrial powers.

France is among the globe's oldest nations, the product of an alliance of duchies and principalities under a single ruler in the Middle Ages. Today, as in that era, central authority is invested in the state, even though a measure of autonomy has been granted to the country's 22 *régions* in recent decades. The French people look to the state as the primary guardian of liberty, and the state in turn provides a generous program of amenities for its citizens, from free education to health care and pension plans. Even so, this centralist tendency is often at odds with another longstanding theme of the French nation: the insistence on the supremacy of the individual. On this matter historian Jules Michelet remarked, "England is an empire, Germany is a nation, a race, France is a person." Statesman Charles de Gaulle, too, famously complained, "Only peril can bring the French together. One can't impose unity out of the blue on a country that has 265 kinds of cheese."

This tendency toward individualism joins with a pluralist outlook and a great interest in the larger world. Even though its imperialist stage was driven by the impulse to civilize that world according to French standards (*la mission civilisatrice*), the French still note approvingly the words of writer Gustave Flaubert:

> I am no more modern than I am ancient, no more French than Chinese; and the idea of *la patrie*, the fatherland—that is, the obligation to live on a bit of earth coloured red or blue on a map, and to detest the other bits coloured green or black—has always seemed to me narrow, restricted, and ferociously stupid.

At once universal and particular, French culture has spread far and greatly influenced the development of art and science, particularly anthropology, philosophy, and sociology.

France has also been influential in government and civil affairs, giving the world important democratic ideals in the age of the Enlightenment and the French Revolution and inspiring the growth of reformist and even revolutionary movements for generations. The present Fifth Republic has, however, enjoyed notable stability since its promulgation on Sept. 28, 1958, marked by a tremendous growth in private initiative and the rise of centrist politics. Although France has engaged in long-running disputes with other European powers (and, from time to time, with the United States, its longtime ally), it emerged as a leading member in the European Union (EU) and its predecessors. From 1966 to 1995 France did not participate in the integrated military structure of the North Atlantic Treaty Organization (NATO), retaining full control over its own air, ground, and naval forces, though since 1995 France has been represented on the NATO Military Committee. As one of the five permanent members of the United Nations Security Council—together with the United States, Russia, the United Kingdom, and China—France has the right to veto decisions put to the council.

The capital and by far the most important city of France is Paris, one of the world's preeminent cultural and commercial centres. A majestic city known as the *ville lumière*, or "city of light," Paris has often been remade, most famously in the mid-19th century under the command of Georges-Eugène, Baron Haussman, who was committed to Napoleon III's vision of a modern city free of the choleric swamps and congested alleys of old, with broad avenues and a regular plan. Paris is now a sprawling metropolis, one of Europe's largest conurbations, but its historic heart can still be traversed in an evening's walk. Confident that their city stood at the very centre of the world, Parisians were once given to referring to their country as having two parts, Paris and *le désert*, the wasteland beyond it. Metropolitan Paris now extends far beyond its ancient suburbs into the countryside, however, and nearly every French town and village now numbers a retiree or two driven from the city by the high cost of living, so that, in a sense, Paris has come to embrace the desert and the desert Paris.

© 2001 Corbis

Mont-Saint-Michel at twilight.

Among France's other major cities are Lyon, located along an ancient Rhône valley trade route linking the North Sea and the Mediterranean; Marseille, a multiethnic port on the Mediterranean founded as an entrepôt for Greek and Carthaginian traders in the 6th century AD; Nantes, an industrial centre and deepwater harbour along the Atlantic coast; and Bordeaux, located in southwestern France along the Garonne River.

This article treats the physical and human geography of France and its history. For discussion of the major cities of France, see the *Macropædia* articles PARIS and MARSEILLE. To locate discussion of the overseas departments of French Guiana, Guadeloupe, Martinique, and Réunion, see the *Index*. For information on Corsica, see the *Micropædia* article CORSICA.

This article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

Land

France lies near the western end of the great Eurasian landmass, largely between latitudes 42° and 51° N. Roughly hexagonal in outline, its continental territory is bordered to the northeast by Belgium and Luxembourg, to the east by Germany, Switzerland, and Italy, to the south by the Mediterranean Sea, Spain, and Andorra, to the west by the Bay of Biscay, and to the northwest by the English Channel (La Manche). To the north, France faces southeastern England across the narrow Strait of Dover (Pas de Calais). Monaco is an independent enclave on the south coast, while the island of Corsica in the Mediterranean is treated as an integral part of the country.

RELIEF

The French landscape, for the most part, is composed of relatively low-lying plains, plateaus, and older mountain blocks, or massifs. This pattern clearly predominates over that of the younger, high ranges, such as the Alps and the Pyrenees. The diversity of the land is typical of Continental Europe.

The three geologic regions

Three main geologic regions are distinguishable: the skeletal remains of ancient mountains that make up the Hercynian massifs; the northern and western plains; and the higher young fold mountains in the south and southeast, including the Alps and the Pyrenees, with their attendant narrow plains. Much of the detailed relief can be attributed geologically to the varying differences in the resistance of rocks to erosion. A great deal of the present landscape detail is due to glaciation during the Pleistocene Epoch (1,600,000 to 10,000 years ago). France lay outside the range of the great ice sheets that descended upon northern Europe, so the direct sculpting of the land by ice was restricted to the Alps, the Pyrenees, the Vosges, Corsica, and the highest summits of the Massif Central. Just outside these glacial areas, in what are known as periglacial lands, repeated freezing and thawing of unprotected surfaces modified slopes by the movement of waste sheets (formed of shattered bedrock), producing very much the landscape that exists today. Pleistocene periglacial action generated the sheets of the fine windblown *limon*, or loess, that is the basis of the most fertile lowland soils, and it possibly also created the Landes, a sandy plain in southwestern France. The development of river terraces (flat, raised surfaces alongside valleys) was another characteristic of periglacial action.

The Hercynian massifs. The physical structure of France is dominated by a group of ancient mountains in the shape of a gigantic V, the sides of which form the two branches of Hercynian folding that took place between 345 and 225 million years ago. The eastern branch comprises the Ardennes, the Vosges, and the eastern part of the Massif Central, while the Hercynian massifs to the west comprise the western part of the Massif Central and the Massif Armoricain.

These highlands are composed of resistant metamorphic, crystalline, and sedimentary rocks from the Paleozoic Era (540 to 245 million years ago), the last including coal deposits. They share the common characteristic of repeated planation, or flattening.

The Ardennes. The Ardennes massif is an extension, from Belgium into France, of the great Rhine Uplands, characterized by rocks of slate and quartz from the Paleozoic Era. Differential erosion of Paleozoic rocks has produced long ridges alternating with open valleys crossed by the Sambre and Meuse rivers.

The Vosges. The Alpine earth movements produced a great upswelling along the line of the present upper Rhine, leaving the Vosges with steep eastern slopes that descend to a rift valley containing the plains of Alsace and Baden; on the west the upland descends rather gently into the scarplands of Lorraine. The Vosges reaches its maximum elevation in the south, near the Alps, where crystalline rocks are exposed; the highest summits are called *ballons*, and the highest is the Ballon de Guebwiller (Mount Guebwiller), with an elevation of 4,669 feet (1,423 metres). To the north the Vosges massif dips beneath a cover of forested sandstone from the Triassic Period (245 to 208 million years ago).

The *ballons*

The Massif Central. The vast plateau of the Massif Central covers about 33,000 square miles (86,000 square km), or some one-sixth of the area of the country. The Massif Central borders the Rhône-Saône valley to the east, the Languedoc lowlands to the south, the Aquitaine Basin to the southwest, and the Paris Basin to the north. Much of the western massif, notably Limousin, consists of monotonous erosion surfaces. The centre and eastern parts of the massif were much fractured in the course of the Alpine movements, leaving behind upthrust blocks, of which the

Cinder cones of the Chaîne des Puys in the Massif Central.
© Christian Kempf from TSW—CLICK/Chicago

most conspicuous is the Morvan, the forested bastion of the northeastern corner of the massif. Downfaulted basins filled with Tertiary sediments (those formed 66.4 to 1.6 million years ago), such as the Limagne near the city of Clermont-Ferrand in south-central France, were also formed. Faulting was associated with volcanic activity, which in the central part of the region formed the vast and complex structures of the massifs of Cantal and Monts Dore, where the Sancy Hill (Puy de Sancy), at 6,184 feet, is the highest summit of the Massif Central. Farther west, on the fringe of the Limagne, is the extraordinary Chaîne des Puys, whose numerous cinder cones were formed only about 10,000 years ago and still retain the newness of their craters, lava flows, and other volcanic features. Numerous mineral springs, such as those at Vichy in the central Auvergne region, are a relic of volcanic activity.

The eastern and southern Massif Central

The eastern and southern portions of the massif, from the Morvan through the Cévennes to the final southwestern termination of the massif in the Noire Mountains (Montagne Noire), are marked by a series of hill masses that overlook the lowlands of the Rhône-Saône river valley and the *région* of Languedoc-Roussillon; at least one of these uplands, Beaujolais, has become famous for the grapevines grown at its foot. Between the hill masses lie infolded coal deposits at locations such as Alès, Decazeville, Saint-Étienne, and Blanzy (Le Creusot) that are of more historical than contemporary importance. To the southwest the rocks of the massif are overlain by a great thickness of limestones (*causses*) from the Jurassic Period (208 to 144 million years ago). Lacking in surface water and little populated, this portion of the massif is crossed by rivers that trench dramatic gorges, notably that of the Tarn. Extensive cave systems bear remains of prehistoric art, such as that of Pêche-Merle in the Lot valley and the Lascaux Grotto in the Vézère valley.

The Massif Armoricain. The Massif Armoricain is contained mostly within the *région* of Brittany (Bretagne), a peninsula washed by the Bay of Biscay on the south and the English Channel on the north. The massif continues beyond Brittany eastward and across the Loire to the south. It is much lower than the other Hercynian massif; its highest point, the Mont des Avaloirs, on the eastern edge of the massif, attains an elevation of 1,368 feet.

The great lowlands. *The Paris Basin.* Between the Ardennes, the Vosges, the Massif Central, and the Massif Armoricain lie the sedimentary beds that make up the Paris Basin. Alternating beds of limestone, sand, and clay dip toward the central Paris Basin, their outcrops forming concentric patterns. Especially to the east, erosion has left the more resistant rocks, usually limestones, with a steep, outward-facing scarp edge and a gentler slope toward the centre of the basin. The central Paris Basin is filled by Tertiary rocks, mostly limestones, that form the level plateaus of regions such as Beauce, Brie, Île-de-France, Valois, and Soissonnais. This area is mostly covered with windblown *limon*, which is the basis of an excellent loamy soil. The limestone levels overlap in sandwich formation. Eroded remnants of higher formations have been left behind as isolated hills called buttes, perhaps the most famous of which is in Paris—the Butte de Montmartre, on which is one of the city's most famous districts. Sandy areas adjoining the limestone formations bear forests, such as the Forest of Fontainebleau, southwest of Paris. In the east, in the regions of Lorraine and Burgundy, are Triassic and Jurassic rocks; among the scarps the Moselle Hills are noted for their *minette*, low-grade iron ore. In the western part of the Paris Basin, scarps in the Jurassic and Cretaceous rocks of Normandy are not prominent. The chalk plateau is trenched by the lower Seine in a course marked by spectacular meanders and river cliffs. Farther north, many stretches of magnificent white chalk cliffs line the English Channel coast.

The chalk plateau

The Flanders Plain. In the extreme north the French boundary includes a small part of the Anglo-Belgian basin. Coastal sand dunes protect the reclaimed marshes of French Flanders from invasion by the sea.

The Alsace Plain. East of the Paris Basin is the Alsace Plain. The terrace and foothills bordering the Rhine are covered with soil-enriching *limon*. Alluvial fans, which are laid down by tributaries emerging from the Vosges, and much of the floodplain of the Rhine and its major tributary, the Ill River, are forested. The Sundgau region of the Alsace Plain, which lies between the Jura and the Ill River above Mulhouse, is another great alluvial fan overlaying impermeable clays, which hold up numerous lakes. The Rhine River and its tributaries continue to deposit thick sediments on the floodplain.

The Loire plains. Toward the southwest the Paris Basin opens on a group of plains that follow the Loire valley. The hills of this area, such as the limestone plateaus of the Touraine region and the crystalline plateaus of the Anjou and Vendée areas, are cut by the broad valleys of the Loire and its tributaries. The middle Loire valley, which varies in width from about 3 to 6 miles (about 5 to 10 kilometres), is famous for its châteaus and its scenic beauty.

The Aquitaine Basin. The Loire countryside links with the Aquitaine Basin of southwestern France through the gap known as the Gate of Poitou. The Aquitaine Basin is much smaller than the Paris Basin, and, while it is bounded in the south by the Pyrenees, in the northeast it runs into the low foothills of the Massif Central.

The younger mountains and adjacent plains. *Pyrenees, Jura, and Alps.* The Pyrenees, whose foothills shelter the picturesque countryside of the Basque region, constitute the most ancient of the more recently formed mountains in France. They stretch for more than 280 miles (450 kilometres), making a natural barrier between France and Spain. Their formation, which began in the Mesozoic Era (245 to 66.4 million years ago), continued in the Tertiary and perhaps even in the beginning of the Quaternary, about 1.6 million years ago. The central and highest part of the barrier is composed of a series of parallel chains with only a few, difficult-to-reach passes that have sheer drops. A section of the mountain chain centring on Mont Perdu (Spanish: Monte Perdido) was named a UNESCO World Heritage site in 1997.

The Jura Mountains, extending into Switzerland, are composed of folded limestone. The northeastern part of the Jura, which has the most pronounced folding, is in Switzerland. The highest point, however, is Mount Neige (5,636 feet), in France.

The French Alps are only a part of the great chain that extends across Europe, but they include its highest point, Mont Blanc (15,771 feet [4,807 metres]). These majestic mountains were formed in a series of foldings that lasted from the beginning of the Tertiary to the Quaternary Peri-

MAP INDEX

Political subdivisions

Regions

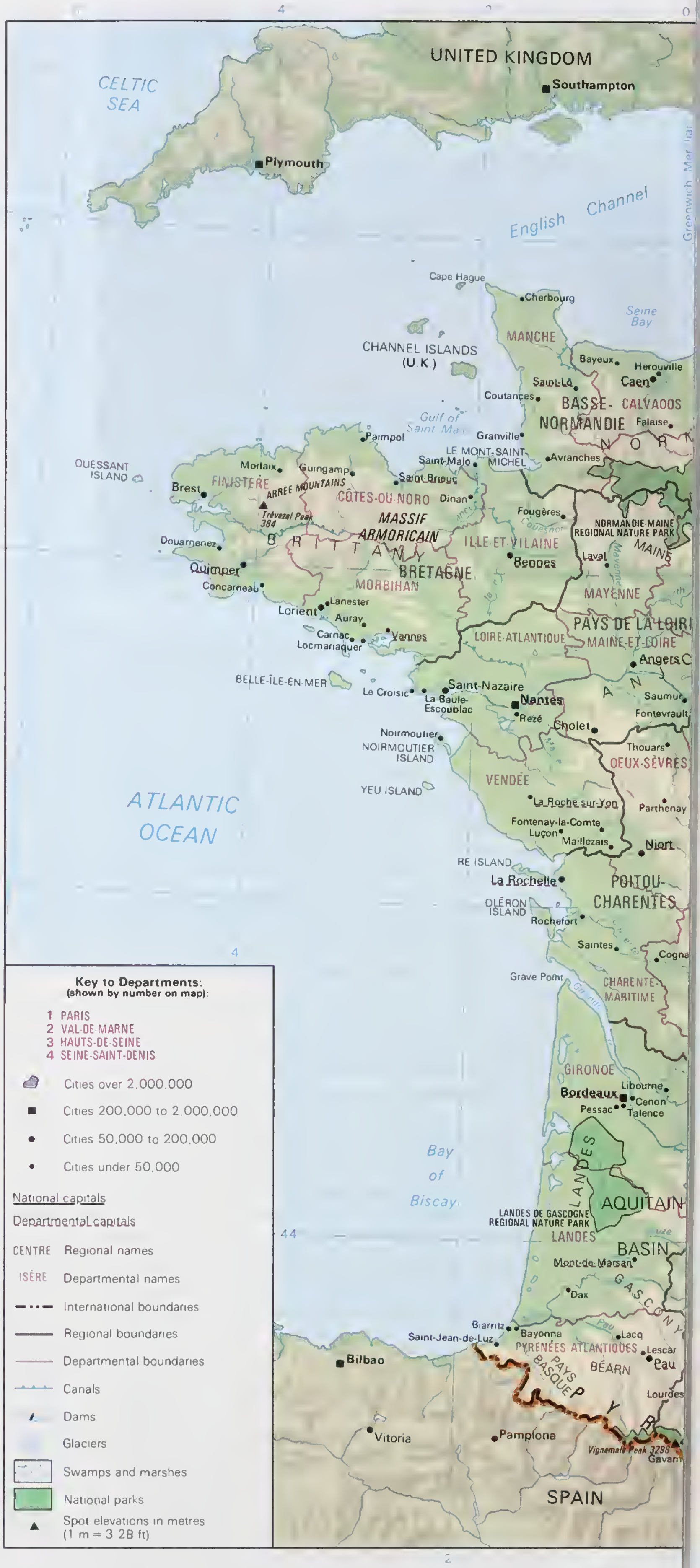
Alsace	48 30 N	7 30 E
Aquitaine	44 35 N	1 00 E
Auvergne	45 20 N	3 00 E
Basse-Normandie	49 00 N	1 00 W
Bourgogne (Burgundy)	47 00 N	4 30 E
Bretagne (Brittany)	48 00 N	3 00 W
Centre	47 30 N	1 50 E
Champagne-Ardenne	48 50 N	4 20 E
Corse	42 00 N	9 00 E
Franche-Comté	47 00 N	6 00 E
Haute-Normandie	49 30 N	1 00 E
Île-de-France	49 00 N	2 20 E
Languedoc-Roussillon	43 30 N	3 30 E
Limousin	45 30 N	1 50 E
Lorraine	49 00 N	6 00 E
Midi-Pyrénées	43 45 N	1 15 E
Nord-Pas-de-Calais	50 20 N	3 40 E
Pays de la Loire	47 30 N	0 45 W
Picardie (Picardy)	50 00 N	3 30 E
Poitou-Charentes	46 00 N	0 00
Provence-Alpes-Côte d'Azur	43 45 N	6 15 E
Rhône-Alpes	45 20 N	5 45 E

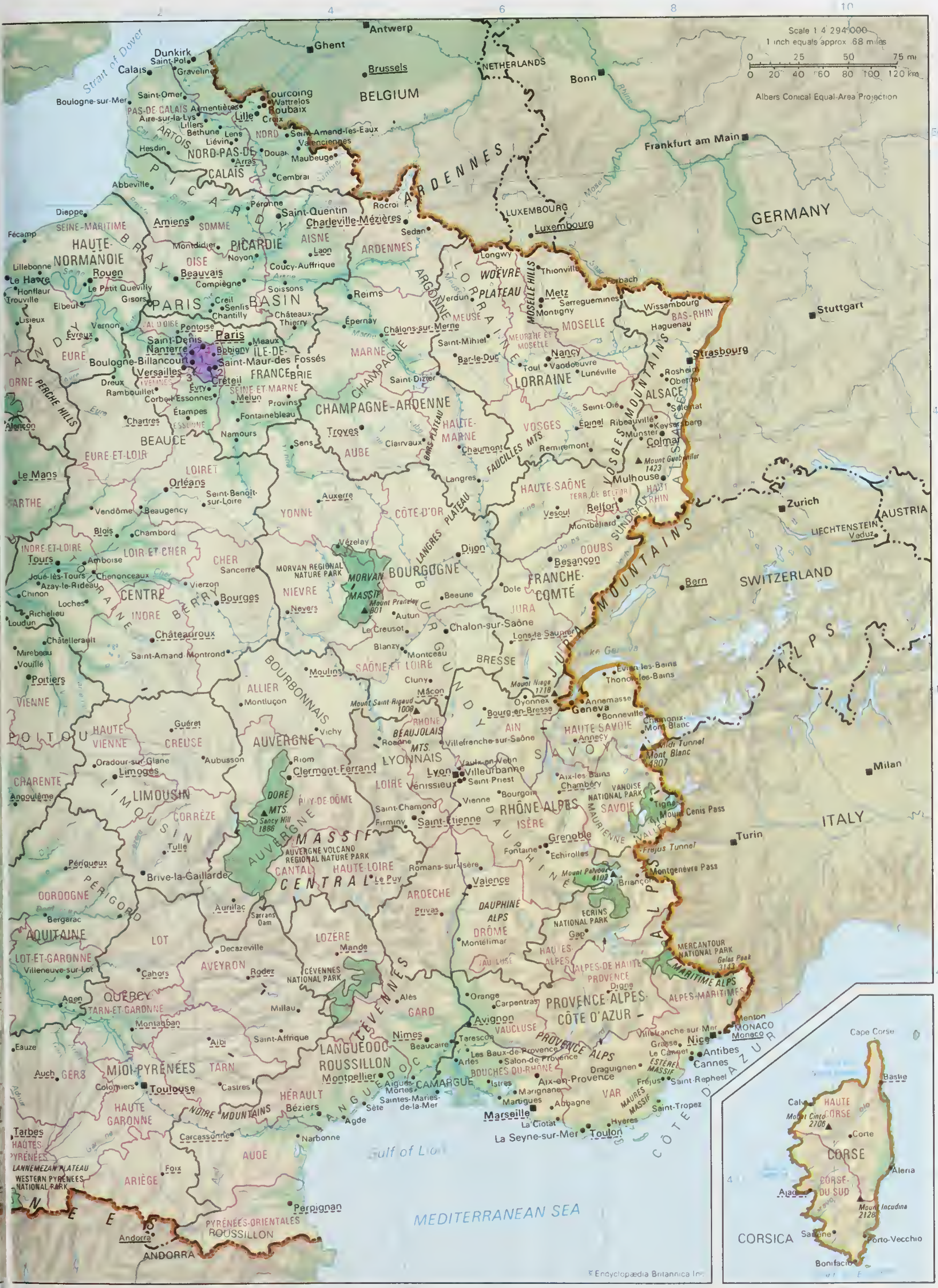
Departments

Ain	46 10 N	5 20 E
Aisne	49 30 N	3 30 E
Allier	46 30 N	3 00 E
Alpes-de-Haute-Provence	44 10 N	6 00 E
Alpes-Maritimes	44 00 N	7 10 E
Ardèche	44 40 N	4 20 E
Ardennes	49 40 N	4 40 E
Ariège	43 00 N	1 30 E
Aube	48 15 N	4 05 E
Aude	43 05 N	2 30 E
Aveyron	44 15 N	2 30 E
Bas-Rhin	48 35 N	7 40 E
Belfort, Territoire de	47 45 N	7 00 E
Bouches-du-Rhône	43 30 N	5 00 E
Calvados	49 10 N	0 30 W
Cantal	45 05 N	2 40 E
Charente	45 40 N	0 05 E
Charente-Maritime	45 30 N	0 45 W
Cher	47 00 N	2 30 E
Corrèze	45 15 N	1 50 E
Corse-du-Sud	41 50 N	9 00 E
Côte-d'Or	47 30 N	4 50 E
Côtes-du-Nord	48 25 N	2 40 W
Creuse	46 05 N	2 00 E
Deux-Sèvres	46 30 N	0 15 W
Dordogne	45 10 N	0 50 E
Doubs	47 10 N	6 25 E
Drôme	44 35 N	5 10 E
Essonne	48 30 N	2 15 E
Eure	49 10 N	1 00 E
Eure-et-Loir	48 30 N	1 30 E
Finistère	48 20 N	4 00 W
Gard	44 00 N	4 00 E
Gers	43 40 N	0 30 E
Gironde	44 55 N	0 30 W
Haut-Rhin	48 00 N	7 20 E
Haute-Corse	42 30 N	9 15 E
Haute-Garonne	43 25 N	1 30 E
Haute-Loire	45 05 N	4 00 E
Haute-Marne	48 05 N	5 10 E
Haute-Saône	47 40 N	6 10 E
Haute-Savoie	46 00 N	6 20 E
Haute-Vienne	45 50 N	1 10 E
Hautes-Alpes	44 40 N	6 30 E
Hautes-Pyrénées	43 00 N	0 10 E
Hauts-de-Seine	48 51 N	2 12 E
Hérault	43 40 N	3 30 E
Ille-et-Vilaine	48 10 N	1 30 W
Indre	46 50 N	1 40 E
Indre-et-Loire	47 15 N	0 45 E
Isère	45 10 N	5 50 E
Jura	46 50 N	5 50 E
Landes	44 00 N	0 50 W
Loir-et-Cher	47 30 N	1 30 E
Loire	45 30 N	4 00 E
Loire-Atlantique	47 15 N	1 50 W
Loiret	47 55 N	2 20 E
Lot	44 30 N	1 30 E
Lot-et-Garonne	44 20 N	0 30 E
Lozère	44 30 N	3 30 E
Maine-et-Loire	47 30 N	0 20 W
Manche	49 00 N	1 10 W
Marne	48 55 N	4 10 E
Mayenne	48 05 N	0 40 W
Meurthe-et-Moselle	48 35 N	6 10 E
Meuse	49 00 N	5 30 E
Morbihan	47 55 N	2 50 W
Moselle	49 00 N	6 30 E
Nièvre	47 05 N	3 30 E
Nord	50 20 N	3 40 E
Oise	49 30 N	2 30 E
Orne	48 40 N	0 05 E
Paris	48 52 N	2 20 E
Pas-de-Calais	50 30 N	2 20 E
Puy-de-Dôme	45 40 N	3 00 E
Pyrénées-Atlantiques	43 15 N	0 50 W
Pyrénées-Orientales	42 30 N	2 20 E
Rhône	46 00 N	4 30 E
Saône-et-Loire	46 40 N	4 30 E
Sarthe	48 00 N	0 05 E
Savoie	45 30 N	6 25 E
Seine-et-Marne	48 30 N	3 00 E
Seine-Maritime	49 45 N	1 00 E
Seine-Saint-Denis	48 55 N	2 30 E
Somme	49 55 N	2 30 E
Tarn	43 50 N	2 00 E
Tarn-et-Garonne	44 00 N	1 10 E
Val-de-Marne	48 47 N	2 27 E
Val-d'Oise	49 04 N	2 04 E
Var	43 30 N	6 20 E
Vaucluse	44 00 N	5 10 E
Vendée	46 40 N	1 20 W
Vienne	46 30 N	0 30 E
Vosges	48 10 N	6 20 E
Yonne	47 55 N	3 45 E
Yvelines	48 47 N	1 52 E

Cities and towns

Abbeville	50 06 N	1 50 E
Agde	43 19 N	3 28 E
Agen	44 12 N	0 38 E
Aigues-Mortes	43 34 N	4 11 E
Aire-sur-la-Lys	50 38 N	2 24 E
Aix-en-Provence	43 32 N	5 26 E
Aix-les-Bains	45 42 N	5 55 E
Ajaccio	41 55 N	8 44 E
Albi	43 56 N	2 09 E
Alençon	48 26 N	0 05 E
Aleria	42 06 N	9 31 E
Alès	44 08 N	4 05 E
Amboise	47 25 N	0 59 E
Amiens	49 54 N	2 18 E
Angers	47 28 N	0 33 W
Angoulême	45 39 N	0 09 E
Annecy	45 54 N	6 07 E
Annemasse	46 12 N	6 15 E
Antibes	43 35 N	7 07 E
Arles	43 40 N	4 38 E
Armentières	50 41 N	2 53 E
Arras	50 17 N	2 47 E
Aubagne	43 17 N	5 34 E
Aubusson	45 57 N	2 10 E
Auch	43 39 N	0 35 E
Auray	47 40 N	2 59 W
Aurillac	44 55 N	2 27 E
Autun	46 57 N	4 18 E
Auxerre	47 48 N	3 34 E
Avignon	43 57 N	4 49 E
Avranches	48 41 N	1 22 W
Azay-le-Rideau	47 16 N	0 28 E
Bar-le-Duc	48 47 N	5 10 E
Bastia	42 42 N	9 27 E
Bayeux	49 16 N	0 42 W
Bayonne	43 29 N	1 29 W
Beaucaire	43 48 N	4 38 E
Beaugency	47 47 N	1 38 E
Beaune	47 02 N	4 50 E




Scale 1:4 294 000
1 inch equals approx. 68 miles
0 25 50 75 mi
0 20 40 60 80 100 120 km
Albers Conical Equal-Area Projection
Strait of Dover
Dunkirk
Calais
Saint-Pol
Gravelines
Boulogne-sur-Mer
Saint-Omer
PAS-DE-CALAIS
Armentières
Aire-sur-la-Lys
Tourcoing
Wattrelos
Roubaix
Lille
Croix
ARTOIS
Lillers
Bethune
Lens
Liévin
Arras
NORD
Saint-Amand-les-Eaux
Valenciennes
Hesdin
NORD-PAS-DE-CALAIS
Douai
Maubeuge
Cambrai
Antwerp
Ghent
Brussels
BELGIUM
NETHERLANDS
Bonn
Frankfurt am Main
ARDENNES
Rocroi
Abbeville
PICARDY
Péronne
Saint-Quentin
Charleville-Mézières
Sedan
LUXEMBOURG
Luxembourg
GERMANY
Dieppe
SEINE-MARITIME
Amiens
SOMME
Fécamp
HAUTE-NORMANDIE
BRAY
Montdidier
PICARDIE
AISNE
Laon
Noyon
Lillebonne
Rouen
OISE
Coucy-Auffrique
Le Havre
Honfleur
Le Petit Quevilly
Beauvais
Compiègne
Trouville
Elbeuf
Gisors
Soissons
PARIS BASIN
Creil
Senlis
Chantilly
Château-Thierry
Reims
Épernay
ARDENNES
ARGONNE
Verdun
LORRAINE PLATEAU
WOËVRE
MOSELLE HILLS
Longwy
Thionville
Metz
Sarreguemines
Forbach
Montigny
MEUSE
MOSELLE
Wissembourg
Haguenau
BAS-RHIN
Stuttgart
Strasbourg
Lisieux
Vernon
Évreux
EURE
VAL-D'OISE
Pontoise
Saint-Denis
Nanterre
Paris
Bobigny
Meaux
Boulogne-Billancourt
ÎLE-DE-FRANCE
Saint-Maur-des Fossés
Versailles
Créteil
BRIE
Dreux
Rambouillet
YVELINES
Évry
Corbeil-Essonnes
SEINE-ET-MARNE
Melun
Étampes
ESSONNE
Fontainebleau
Provins
Châlons-sur-Marne
MARNE
CHAMPAGNE
Saint-Dizier
Saint-Mihiel
Bar-le-Duc
MEURTHE-ET-MOSELLE
Toul
Nancy
Vandoeuvre
Lunéville
LORRAINE
Rosheim
Obernai
ALSACE
VOSGES MOUNTAINS
Saint-Dié
Sélestat
Épinal
Ribeauvillé
Keysersberg
Munster
Colmar
ORNE
PERCHE HILLS
Alençon
Chartres
EURE-ET-LOIR
BEAUCE
Nemours
CHAMPAGNE-ARDENNE
HAUTE-MARNE
Troyes
AUBE
Sens
Clairvaux
BARS PLATEAU
Chaumont
VOSGES
FAUCILLES MTS
Remiremont
Mount Guebwiller 1423
Mulhouse
HAUT-RHIN
Le Mans
SARTHE
LOIRET
Orléans
Saint-Benoit-sur-Loire
Vendôme
Beaugency
Blois
Chambord
YONNE
Auxerre
Langres
CÔTE-D'OR
PLATEAU
HAUTE-SAÔNE
TERR. DE BELFORT
Vesoul
Belfort
Montbéliard
SUNDGAU
Zurich
AUSTRIA
LIECHTENSTEIN
Vaduz
INDRE-ET-LOIRE
Tours
Amboise
LOIR-ET-CHER
Joué-lès-Tours
Chenonceaux
Azay-le-Rideau
Chinon
CHER
Sancerre
Vézelay
MORVAN REGIONAL NATURE PARK
MORVAN MASSIF
Mount Preneley 901
LANGRES PLATEAU
Dijon
BOURGOGNE
DOUBS
Besançon
FRANCHE-COMTÉ
Dole
JURA
Bern
SWITZERLAND
Loches
CENTRE
Vierzon
Bourges
NIEVRE
Nevers
Autun
Beaune
Le Creusot
Chalon-sur-Saône
Lons-le-Saunier
JURA MOUNTAINS
Richelieu
Loudun
INDRE
BERRY
TOURAINE
Châteauroux
Châtellerault
Mirebeau
Vouillé
Saint-Amand-Montrond
Poitiers
VIENNE
BOURBONNAIS
Blanzy
Montceau
SAÔNE-ET-LOIRE
BRESSE
Moulins
ALLIER
Montluçon
Cluny
Mâcon
Mount Saint-Rigaud 1009
Mount Niege 1718
Oyonnax
Bourg-en-Bresse
Lake Geneva
Évian-les-Bains
Thonon-les-Bains
Annemasse
Geneva
Bonneville
ALPS
BURGUNDY
POITOU
HAUTE-VIENNE
Guéret
CREUSE
AUVERGNE
Vichy
RHÔNE
BEAUJOLAIS
Roanne
MTS.
Villefranche-sur-Saône
AIN
HAUTE-SAVOIE
Annecy
Chamonix-Mont Blanc
Midi Tunnel
Mont Blanc 4807
SAVOY
Milan
CHARENTE
Angoulême
Oradour-sur-Glane
Limoges
Aubusson
Riom
Clermont-Ferrand
LYONNAIS
Lyon
Vaulx-en-Velin
Villeurbanne
Vénissieux
Saint-Priest
LOIRE
Aix-les-Bains
Chambéry
VANOISE NATIONAL PARK
SAVOIE
MAURIENNE VALLEY
Tignes
Mount Cenis Pass
ITALY
LIMOUSIN
LIMOUSIN
DORE MTS.
Sancy Hill 1886
PUY-DE-DÔME
Saint-Chamond
Firminy
Saint-Étienne
Vienne
Bourgoin
RHÔNE-ALPES
ISÈRE
DAUPHINÉ
Grenoble
Fontaine
Échirolles
Fréjus Tunnel
Turin
CORRÈZE
Tulle
MASSIF CENTRAL
AUVERGNE VOLCANO REGIONAL NATURE PARK
AUVERGNE
CANTAL
HAUTE-LOIRE
Le Puy
Romans-sur-Isère
Valence
Mount Pelvoux 4103
Briançon
Montgenèvre Pass
Périgueux
Brive-la-Gaillarde
PÉRIGORD
DORDOGNE
Bergerac
Aurillac
Sarrans Dam
ARDECHE
Privas
DAUPHINE ALPS
DRÔME
Montélimar
ÉCRINS NATIONAL PARK
Gap
HAUTES-ALPES
AQUITAINE
LOT
LOT-ET-GARONNE
Villeneuve-sur-Lot
Decazeville
AVEYRON
Rodez
LOZÈRE
Mende
CÉVENNES NATIONAL PARK
CÉVENNES
VAUCLUSE
ALPES-DE-HAUTE-PROVENCE
Digne
MERCANTOUR NATIONAL PARK
Gelas Peak 3143
MARITIME ALPS
ALPES-MARITIMES
Cahors
Agen
QUERCY
TARN-ET-GARONNE
Millau
Alès
GARD
Orange
Carpentras
PROVENCE-ALPES-CÔTE D'AZUR
Montauban
Albi
Saint-Affrique
Avignon
Nîmes
Beaucaire
Tarascon
Les Baux-de-Provence
Salon-de-Provence
Arles
PROVENCE ALPS
Grasse
Nice
Menton
MONACO
Monaco
Villefranche-sur-Mer
Le Cannet
Antibes
Cannes
ESTÉREL MASSIF
Draguignan
Fréjus
Saint-Raphael
Eauze
GERS
Auch
MIDI-PYRÉNÉES
TARN
Colomiers
Toulouse
Castres
LANGUEDOC-ROUSSILLON
Montpellier
Aigues-Mortes
CAMARGUE
Saintes-Maries-de-la-Mer
BOUCHES-DU-RHÔNE
Istres
Aix-en-Provence
Marignane
Martigues
Aubagne
VAR
MAURES MASSIF
Saint-Tropez
CÔTE D'AZUR
HÉRAULT
Béziers
Sète
Agde
LANGUEDOC
Marseille
La Ciotat
La Seyne-sur-Mer
Toulon
Hyères
HAUTE-GARONNE
NOIRE MOUNTAINS
Tarbes
HAUTES-PYRÉNÉES
LANNEMEZAN PLATEAU
WESTERN PYRENEES NATIONAL PARK
Carcassonne
Narbonne
AUDE
Gulf of Lion
Foix
ARIÈGE
PYRENEES
Perpignan
PYRÉNÉES-ORIENTALES
ROUSSILLON
MEDITERRANEAN SEA
Andorra
ANDORRA
Cape Corse
Bastia
HAUTE-CORSE
Calvi
Mount Cinto 2706
Corte
CORSE
Aleria
CORSE-DU-SUD
Ajaccio
Mount Incudine 2128
CORSICA
Sartène
Porto-Vecchio
Bonifacio
© Encyclopædia Britannica Inc.
2
4
6
8
10
50°
48
46
44
42

Beauvais	49 26 N	2 05 E
Belfort	47 38 N	6 52 E
Bergerac	44 51 N	0 29 E
Besançon	47 15 N	6 02 E
Béthune	50 32 N	2 38 E
Béziers	43 21 N	3 15 E
Biarritz	43 29 N	1 34 W
Blanzy	46 42 N	4 23 E
Blois	47 35 N	1 20 E
Bobigny	48 54 N	2 27 E
Bonifacio	41 23 N	9 09 E
Bonneville	46 05 N	6 25 E
Bordeaux	44 50 N	0 34 W
Boulogne-Billancourt	48 50 N	2 15 E
Boulogne-sur-Mer	50 43 N	1 37 E
Bourg-en-Bresse	46 12 N	5 13 E
Bourges	47 05 N	2 24 E
Bourgoin	45 35 N	5 17 E
Brest	48 24 N	4 29 W
Briançon	44 54 N	6 39 E
Brive-la-Gaillarde	45 09 N	1 32 E
Caen	49 11 N	0 21 W
Cahors	44 26 N	1 26 E
Calais	50 57 N	1 50 E
Calvi	42 34 N	8 45 E
Cambrai	50 10 N	3 14 E
Cannes	43 33 N	7 01 E
Carcassonne	43 13 N	2 21 E
Carnac	47 35 N	3 05 W
Carpentras	44 03 N	5 03 E
Castres	43 36 N	2 15 E
Cenon	44 51 N	0 32 W
Chalon-sur-Saône	46 47 N	4 51 E
Châlons-sur-Marne	48 57 N	4 22 E
Chambéry	45 34 N	5 56 E
Chambord	47 37 N	1 31 E
Chamonix-Mont-Blanc	45 55 N	6 52 E
Chantilly	49 12 N	2 28 E
Charleville-Mézières	49 46 N	4 43 E
Chartres	48 27 N	1 30 E
Châteauroux	46 49 N	1 42 E
Châteaux-Thierry	49 03 N	3 24 E
Châtellerault	46 48 N	0 32 E
Chaumont	48 07 N	5 08 E
Chenonceaux	47 20 N	1 04 E
Cherbourg	49 39 N	1 39 W
Chinon	47 10 N	0 15 E
Cholet	47 04 N	0 53 W
Clairvaux	48 09 N	4 47 E
Clermont-Ferrand	45 47 N	3 05 E
Cluny	46 26 N	4 39 E
Cognac	45 42 N	0 20 W
Colmar	48 05 N	7 22 E
Colomiers	43 37 N	1 21 E
Compiègne	49 25 N	2 50 E
Concarneau	47 52 N	3 55 W
Corbeil-Essonnes	48 36 N	2 29 E
Corte	42 18 N	9 09 E
Coucy-Auffrique	49 31 N	3 19 E
Coutances	49 03 N	1 26 W
Creil	49 16 N	2 29 E
Créteil	48 47 N	2 28 E
Croix	50 40 N	3 09 E
Dax	43 43 N	1 03 W
Decazeville	44 33 N	2 15 E
Dieppe	49 56 N	1 05 E
Digne	44 06 N	6 14 E
Dijon	47 19 N	5 01 E
Dinan	48 27 N	2 02 W
Dole	47 06 N	5 30 E
Douai	50 22 N	3 04 E
Douarnenez	48 06 N	4 20 W
Draguignan	43 32 N	6 28 E
Dreux	48 44 N	1 22 E
Dunkirk (Dunkerque)	51 03 N	2 22 E
Eauze	43 52 N	0 06 E
Échirolles	45 08 N	5 43 E
Elbeuf	49 17 N	1 00 E
Épernay	49 03 N	3 57 E
Épinal	48 11 N	6 27 E
Étampes	48 26 N	2 09 E
Évian-les-Bains	46 23 N	6 35 E
Évreux	49 01 N	1 09 E
Évry	48 38 N	2 27 E
Falaise	48 54 N	0 12 W
Fécamp	49 45 N	0 22 E
Firminy	45 23 N	4 18 E
Foix	42 58 N	1 36 E
Fontaine	45 11 N	5 40 E
Fontainebleau	48 24 N	2 42 E
Fontenay-le-Comte	46 28 N	0 49 W
Fontevrault	47 11 N	0 03 E
Forbach	49 11 N	6 54 E
Fougères	48 21 N	1 12 W
Fréjus	43 26 N	6 44 E
Gap	44 34 N	6 05 E
Gavarnie	42 44 N	0 00
Gisors	49 17 N	1 47 E
Granville	48 50 N	1 36 W
Grasse	43 40 N	6 55 E
Gravelines	50 59 N	2 07 E
Grenoble	45 10 N	5 43 E
Guéret	46 10 N	1 52 E
Guingamp	48 33 N	3 09 W
Haguenau	48 49 N	7 47 E
Herouville	49 12 N	0 19 W
Hesdin	50 22 N	2 02 E
Honfleur	49 25 N	0 14 E
Hyères	43 07 N	6 07 E
Istres	43 31 N	4 59 E
Joué-lès-Tours	47 21 N	0 40 E
Kaysersberg	48 08 N	7 15 E
La Baule-Escoublac	47 17 N	2 24 W
La Ciotat	43 10 N	5 36 E
La Rochelle	46 10 N	1 09 W
La Roche-sur-Yon	46 40 N	1 26 W
La Seyne-sur-Mer	43 06 N	5 53 E
Lacq	43 25 N	0 38 W
Lanester	47 45 N	3 21 W
Langres	47 52 N	5 20 E
Laon	49 34 N	3 37 E
Laval	48 04 N	0 46 W
Le Cannet	43 34 N	7 01 E
Le Creusot	46 48 N	4 26 E
Le Croisic	47 18 N	2 31 W
Le Havre	49 30 N	0 08 E
Le Mans	48 00 N	0 12 E
Le Petit Quevilly	49 26 N	1 02 E
Le Puy	45 02 N	3 53 E
Lens	50 26 N	2 50 E
Les Baux-de-Provence	43 45 N	4 48 E
Lescar	43 20 N	0 25 W
Libourne	44 55 N	0 14 W
Liévin	50 25 N	2 46 E
Lille	50 38 N	3 04 E
Lillebonne	49 31 N	0 33 E
Lillers	50 34 N	2 29 E
Limoges	45 45 N	1 20 E
Lisieux	49 09 N	0 14 E
Loches	47 08 N	1 00 E
Locmariaquer	47 34 N	2 57 W
Longwy	49 31 N	5 46 E
Lons-le-Saunier	46 40 N	5 33 E
Lorient	47 45 N	3 22 W
Loudun	47 00 N	0 04 E
Lourdes	43 06 N	0 03 W
Luçon	46 27 N	1 10 W
Lunéville	48 36 N	6 30 E
Lyon	45 45 N	4 51 E
Mâcon	46 18 N	4 50 E
Maillezais	46 22 N	0 44 W
Marignane	43 25 N	5 13 E
Marseille	43 18 N	5 24 E
Martigues	43 24 N	5 03 E
Maubeuge	50 17 N	3 58 E
Meaux	48 57 N	2 52 E
Melun	48 32 N	2 40 E
Mende	44 31 N	3 30 E
Menton	43 47 N	7 30 E
Metz	49 08 N	6 10 E
Millau	44 06 N	3 05 E
Mirebeau	46 47 N	0 11 E
Mont-de-Marsan	43 53 N	0 30 W
Montauban	44 01 N	1 21 E
Montbéliard	47 31 N	6 48 E
Montceau	46 40 N	4 22 E
Montdidier	49 39 N	2 34 E
Montélimar	44 34 N	4 45 E
Montigny	49 06 N	6 09 E
Montluçon	46 20 N	2 36 E
Montpellier	43 36 N	3 53 E
Morlaix	48 35 N	3 50 W
Moulins	46 34 N	3 20 E
Mulhouse	47 45 N	7 20 E
Munster	48 03 N	7 08 E
Nancy	48 41 N	6 12 E
Nanterre	48 54 N	2 12 E
Nantes	47 13 N	1 33 W
Narbonne	43 11 N	3 00 E
Nemours	48 16 N	2 42 E
Nevers	46 59 N	3 10 E
Nice	43 42 N	7 15 E
Nîmes	43 50 N	4 21 E
Niort	46 19 N	0 28 W
Noirmoutier	47 00 N	2 15 W
Noyon	49 35 N	3 00 E
Obernai	48 28 N	7 29 E
Oradour-sur-Glane	45 56 N	1 02 E
Orange	44 08 N	4 48 E
Orléans	47 55 N	1 54 E
Oyonnax	46 15 N	5 40 E
Paimpol	48 46 N	3 03 W
Paris	48 52 N	2 20 E
Parthenay	46 39 N	0 15 W
Pau	43 18 N	0 22 W
Périgueux	45 11 N	0 43 E
Péronne	49 56 N	2 56 E
Perpignan	42 41 N	2 53 E
Pessac	44 48 N	0 37 W
Poitiers	46 35 N	0 20 E
Pontoise	49 03 N	2 06 E
Porto-Vecchio	41 35 N	9 17 E
Privas	44 44 N	4 36 E
Provins	48 33 N	3 18 E
Quimper	48 00 N	4 06 W
Rambouillet	48 39 N	1 50 E
Reims	49 15 N	4 02 E
Remiremont	48 01 N	6 35 E
Rennes	48 05 N	1 41 W
Rezé	47 12 N	1 34 W
Ribeauvillé	48 12 N	7 19 E
Richelieu	47 01 N	0 19 E
Riom	45 54 N	3 07 E
Roanne	46 02 N	4 04 E
Rochefort	45 56 N	0 59 W
Rocroi	49 55 N	4 31 E
Rodez	44 20 N	2 34 E
Romans-sur-Isère	45 03 N	5 03 E
Rosheim	48 30 N	7 28 E
Roubaix	50 42 N	3 10 E
Rouen	49 26 N	1 05 E
Saint-Affrique	43 57 N	2 53 E
Saint-Amand-les-Eaux	50 26 N	3 26 E
Saint-Amand-Montrond	46 43 N	2 31 E
Saint-Benoît-sur-Loire	47 49 N	2 18 E
Saint-Brieuc	48 31 N	2 47 W
Saint-Chamond	45 28 N	4 30 E
Saint-Denis	48 56 N	2 22 E
Saint-Dié	48 17 N	6 57 E
Saint-Dizier	48 38 N	4 57 E
Saint-Étienne	45 26 N	4 24 E
Saint-Jean-de-Luz	43 23 N	1 40 W
Saint-Lô	49 07 N	1 05 W
Saint-Malo	48 39 N	2 01 W
Saint-Maur-des-Fossés	48 48 N	2 30 E
Saint-Mihiel	48 54 N	5 33 E
Saint-Nazaire	47 17 N	2 12 W
Saint-Omer	50 45 N	2 15 E
Saint-Pol	51 02 N	2 21 E
Saint-Priest	45 42 N	4 57 E
Saint-Quentin	49 51 N	3 17 E
Saint-Raphaël	43 25 N	6 46 E
Saint-Tropez	43 16 N	6 38 E
Saintes	45 45 N	0 38 W
Saintes-Maries-de-la-Mer	43 27 N	4 26 E
Salon-de-Provence	43 38 N	5 06 E
Sancerre	47 20 N	2 50 E
Sarreguemines	49 06 N	7 03 E
Sartène	41 37 N	8 59 E
Saumur	47 16 N	0 05 W
Sedan	49 42 N	4 57 E
Sélestat	48 16 N	7 27 E
Senlis	49 12 N	2 35 E
Sens	48 12 N	3 17 E
Sète	43 24 N	3 41 E
Soissons	49 22 N	3 20 E
Strasbourg	48 35 N	7 45 E
Talence	44 49 N	0 36 W
Tarascon	43 48 N	4 40 E
Tarbes	43 14 N	0 05 E
Thionville	49 22 N	6 10 E
Thonon-les-Bains	46 22 N	6 29 E
Thouars	46 58 N	0 13 W
Tignes	45 30 N	6 55 E
Toul	48 41 N	5 54 E
Toulon	43 07 N	5 56 E
Toulouse	43 36 N	1 26 E
Tourcoing	50 43 N	3 09 E
Tours	47 23 N	0 41 E
Trouville	49 22 N	0 05 E
Troyes	48 18 N	4 05 E
Tulle	45 16 N	1 46 E
Valence	44 56 N	4 54 E
Valenciennes	50 21 N	3 32 E
Vandoeuvre	48 39 N	6 11 E
Vannes	47 40 N	2 45 W
Vaulx-en-Velin	45 47 N	4 56 E
Vendôme	47 48 N	1 04 E
Vénissieux	45 41 N	4 53 E
Verdun	49 10 N	5 23 E
Vernon	49 05 N	1 29 E
Versailles	48 48 N	2 08 E
Vesoul	47 38 N	6 10 E
Vézelay	47 28 N	3 44 E
Vichy	46 07 N	3 25 E
Vienne	45 31 N	4 52 E
Vierzon	47 13 N	2 05 E
Villefranche-sur-Mer	43 42 N	7 19 E
Villefranche-sur-Saône	45 59 N	4 43 E
Villeneuve-sur-Lot	44 24 N	0 43 E
Villeurbanne	45 46 N	4 53 E
Vouillé	46 38 N	0 10 E
Wattrelos	50 42 N	3 13 E
Wissembourg	49 02 N	7 57 E

Physical features and points of interest

Aa, *river*	50 44 N	2 18 E
Adour, *river*	43 32 N	1 32 W
Ain, *river*	45 48 N	5 10 E
Aisne, *river*	49 26 N	2 50 E
Allier, *river*	46 57 N	3 05 E
Alps, *mountains*	46 25 N	10 00 E
Alsace, *historical region*	48 30 N	7 30 E
Anjou, *historical region*	47 20 N	0 30 W
Annecy, Lake	45 51 N	6 11 E
Aquitaine Basin	44 00 N	0 00
Ardennes, *plateau*	50 10 N	5 30 E
Argonne, *physical region*	49 30 N	5 00 E
Armorican, Massif	48 00 N	3 00 W
Arrats, *river*	44 06 N	0 52 E
Arrée Mountains	48 26 N	3 55 W
Artois, *historical region*	50 30 N	2 30 E
Atlantic Ocean	46 00 N	4 00 W
Aube, *river*	48 34 N	3 43 E
Aude, *river*	43 13 N	3 14 E
Auvergne, *historical region*	45 20 N	3 00 E
Auvergne Volcano National Regional Nature Park	45 30 N	2 50 E
Avre, *river*	48 47 N	1 22 E
Bars Plateau	48 10 N	4 45 E
Basque, *historical region*	43 15 N	1 00 W
Béarn, *historical region*	43 20 N	0 45 W
Beauce, *physical region*	48 22 N	1 50 E
Beaujolais Mountains	46 00 N	4 22 E
Belle-Île-en-Mer, *island*	47 19 N	3 11 W
Berry, *historical region*	47 00 N	2 00 E
Biscay, Bay of	44 00 N	2 00 W
Blanc, Mont	45 50 N	6 52 E
Bonifacio, Strait of	41 18 N	9 15 E
Bourbonnais, *historical region*	46 30 N	3 00 E
Bourget, Lake	45 44 N	5 52 E

Bourgogne, see Burgundy		
Bray, *physical region*	49 46 N	1 26 E
Bresle, *river*	50 04 N	1 22 E
Bresse, *physical region*	46 30 N	5 15 E
Brie, *physical region*	48 40 N	3 20 E
Brittany (Bretagne), *historical region*	48 00 N	3 00 W
Burgundy (Bourgogne), *historical region*	47 00 N	4 30 E
Camargue, *physical region*	43 34 N	4 34 E
Canche, *river*	50 31 N	1 39 E
Central, Massif	45 00 N	3 00 E
Cère, *river*	44 55 N	1 49 E
Cévennes, *mountains*	44 00 N	3 30 E
Cévennes National Park	44 20 N	3 50 E
Champagne, *historical region*	9 00 N	4 30 E
Charente, *river*	45 57 N	1 05 W
Cher, *river*	47 21 N	0 29 E
Cinto, Mount	42 23 N	8 56 E
Corse, Cape	43 00 N	9 23 E
Corsica, *island*	42 00 N	9 00 E
Côte d'Azur	43 30 N	7 00 E
Couesnon, *river*	48 37 N	1 31 W
Creuse, *river*	47 00 N	0 34 E
Cure, *river*	47 40 N	3 41 E
Dauphiné, *historical region*	44 50 N	6 00 E
Dauphiné Alps	44 40 N	6 07 E
Dordogne, *river*	45 02 N	0 35 W
Dore Mountains	45 35 N	2 50 E
Doubs, *river*	46 54 N	5 02 E
Douze, *river*	43 54 N	0 30 W
Dover, Strait of	51 00 N	1 30 E
Drac, *river*	45 13 N	5 41 E
Drôme, *river*	44 46 N	4 46 E
Dronne, *river*	45 02 N	0 09 W
Durance, *river*	43 55 N	4 44 E
Ecrins National Park	44 55 N	6 15 E
English Channel	50 00 N	1 00 W
Epte, *river*	49 04 N	1 31 E
Escaut (Schelde), *river*	51 22 N	4 15 E
Estérel Massif	43 32 N	6 49 E
Eure, *river*	49 18 N	1 12 E
Faucilles Mountains	48 07 N	6 16 E
Fréjus Tunnel	45 07 N	6 40 E
Garonne, *river*	45 02 N	0 36 W
Gartempe, *river*	46 47 N	0 50 E
Gascony (Gascogne), *historical region*	43 30 N	0 00
Gelas Peak	44 07 N	7 23 E
Geneva, Lake	46 25 N	6 30 E
Gers, *river*	44 09 N	0 39 E
Gironde, *estuary*	45 35 N	1 03 W
Golo, *river*	42 31 N	9 32 E
Grave Point	45 34 N	1 04 W
Guebwiller, Mount	47 55 N	7 08 E
Hague, Cape	49 43 N	1 57 W
Ill, *river*	48 40 N	7 53 E
Incudine, Mount	41 51 N	9 12 E
Indre, *river*	47 14 N	0 11 E
Isère, *river*	44 59 N	4 51 E
Isle, *river*	44 55 N	0 15 W
Jura Mountains	47 10 N	7 00 E
Landes, *physical region*	44 15 N	1 00 W
Landes de Gascogne Regional Nature Park	44 30 N	1 00 W
Langres Plateau	47 41 N	5 03 E
Languedoc, *historical region*	44 00 N	4 00 E
Lannemezan Plateau	43 09 N	0 27 E
Limousin, *historical region*	45 30 N	1 50 E
Lion, Gulf of	43 00 N	4 00 E
Loir, *river*	47 33 N	0 32 W
Loire, *river*	47 16 N	2 11 W
Lorraine, *historical region*	49 00 N	6 00 E
Lot, *river*	44 18 N	0 20 E
Lyonnais, *historical region*	45 45 N	4 30 E
Maas, see Meuse		
Maine, *historical region*	48 15 N	0 00
Maine, *river*	47 09 N	1 27 W
Maritime Alps	44 15 N	7 10 E
Marne, *river*	48 49 N	2 24 E
Maures Massif	43 16 N	6 23 E
Maurienne Valley	45 13 N	6 30 E
Mayenne, *river*	47 30 N	0 32 W
Mediterranean Sea	35 00 N	5 00 E
Mercantour National Park	44 15 N	6 48 E
Mer de Glace, *glacier*	45 55 N	6 55 E
Meuse (Maas), *river*	51 49 N	5 01 E
Midi Tunnel	45 52 N	6 53 E
Mont-Saint-Michel, *islet*	48 38 N	1 32 E
Montgenèvre Pass	44 56 N	6 44 E
Morvan Massif	47 05 N	4 00 E
Morvan Regional Nature Park	47 20 N	4 00 E
Moselle, *river*	50 22 N	7 36 E
Moselle Hills	49 30 N	5 45 E
Mount Cenis Pass	45 15 N	6 54 E
Neige, Mount	46 16 N	5 56 E
Noire Mountains	43 28 N	2 18 E
Noirmoutier Island	46 58 N	2 12 W
Normandie-Maine Regional Nature Park	48 35 N	0 30 W
Normandy (Normandie), *historical region*	49 00 N	0 00
Oise, *river*	49 00 N	2 04 E
Oléron Island	45 56 N	1 18 W
Omain, *river*	48 46 N	4 47 E
Orne, *river*	49 18 N	0 15 W
Ouessant Island	48 28 N	5 05 W
Paris Basin	49 00 N	2 00 E
Pau, *river*	43 33 N	1 12 W
Pelvoux, Mount	44 54 N	6 24 E
Perche Hills	48 25 N	0 40 E
Périgord, *historical region*	45 00 N	0 30 E
Picardy (Picardie), *historical region*	50 00 N	3 30 E
Poitou, *historical region*	46 40 N	0 30 W
Preneley, Mount	46 57 N	4 00 E
Provence Alps	43 50 N	6 15 E
Pyrenees, *mountains*	42 40 N	1 00 E
Quercy, *historical region*	4 15 N	1 15 E
Rance, *river*	48 31 N	1 59 W
Re Island	46 12 N	1 25 W
Rhine, *river*	50 00 N	7 55 E
Rhône, *river*	43 20 N	4 50 E
Roussillon, *historical region*	42 30 N	2 30 E
Saar, *river*	49 42 N	6 34 E
Sagone, Gulf of	42 02 N	8 40 E
Saint-Florent, Gulf of	42 45 N	9 16 E
Saint-Malo, Gulf of	48 45 N	2 00 W
Saint-Rigaud, Mount	46 12 N	4 29 E
Sambre, *river*	50 28 N	4 52 E
Sancy Hill	45 32 N	2 50 E
Saône, *river*	45 44 N	4 50 E
Sarrans Dam	44 50 N	2 45 E
Sarthe, *river*	47 30 N	0 32 W
Savoy (Savoie), *historical region*	45 24 N	6 30 E
Schelde, see Escaut		
Seine, *river*	49 26 N	0 26 E
Seine Bay	49 30 N	0 30 W
Serre-Ponçon Reservoir	44 30 N	6 20 E
Somme, *river*	50 11 N	1 39 E
Sundgau, *physical region*	47 40 N	7 15 E
Taravo, *river*	41 42 N	8 48 E
Tarn, *river*	44 06 N	1 02 E
Têt, *river*	42 44 N	3 02 E
Tinée, *river*	43 55 N	7 11 E
Touraine, *historical region*	47 12 N	1 30 E
Trévezel Peak	48 24 N	3 55 W
Vanoise National Park	45 30 N	6 55 E
Verdon, *river*	43 43 N	5 46 E
Vézère, *river*	44 53 N	0 53 E
Vienne, *river*	47 13 N	0 05 E
Vignemale Peak	42 46 N	0 09 W
Vilaine, *river*	47 30 N	2 27 W
Vosges Mountains	48 30 N	7 10 E
Western Pyrenees National Park	42 45 N	0 00
Woëvre Plateau	49 27 N	5 17 E
Yeu Island	46 43 N	2 20 W
Yonne, *river*	48 23 N	2 58 E

od. They include the two greatest regions of permanent snow and glaciers in Europe. The northern Alps are relatively easy to cross because of the numerous valleys created by the movement of glaciers. The relief of the southern Alps is much less orderly, and the valleys, which were not affected by glaciation, form narrow and winding gorges. Like the Pyrenees, the Alps form a natural barrier, dropping sharply down to the Po River plain in Italy.

The French Riviera

The southern plains. Between these young mountains and the ancient Massif Central is a series of plains, including those of the Saône and the Rhône rivers, which extend southward to the great triangular delta of the Rhône on the Mediterranean coast. Its seaward face, the Camargue region, comprises a series of lakes, marshes, and sand spits and includes one of Europe's important wetland nature reserves. West of the Rhône delta the Languedoc coastal plain is broad and rather featureless. At the southwestern end the foothills of the Pyrenees reach to the rocky coast of the Roussillon region. East of the Rhône delta the lowlands are more fragmentary; in the Côte d'Azur region the Alpine foothills and the ancient Maures and Esterel massifs reach to the Mediterranean, forming the coves, capes, and harbours of the country's most famous tourist and retirement area, the French Riviera. Corsica is also highly regarded for its natural scenery. A number of the island's peaks reach over 6,500 feet, with parts of it under wild forest or covered with undergrowth, or maquis.

DRAINAGE

The river systems of France are determined by a major divide in the far eastern part of the country, running from the southern end of the Vosges down the eastern and southeastern edge of the Massif Central to the Noire Mountains, the southwestern promontory of the massif. This divide is broken by occasional cols (depressions) and lowland corridors, notably the Langres Plateau, across the Jurassic outer rim of the Paris Basin. Along the divide originate most of the rivers of the larger, western part of the country, including the Seine and the Loire. Other major rivers include the Garonne, originating in the Pyrenees, and the Rhône and the Rhine, originating in the Alps.

The Seine system. The main river of the Paris Basin, the Seine, 485 miles in length, is joined upstream on the left bank by its tributary the Yonne, on the right bank south of Paris by the Marne, and north of the city by the Oise. While the Seine has a regular flow throughout the year, there may be flooding in the spring and, occasionally more severely, during the customary fall-winter peak of lowland rivers. Efforts have been made to reduce flooding on the Seine and its tributaries by the building of reservoirs. A number of islands dot the Seine along its meandering, generally westward course across the central Paris Basin and through the capital city itself. One of these, the Île de la Cité, forms the very heart of the city of Paris. Eventually the river enters the English Channel at Le Havre.

The Loire system. The Loire, the longest French river, flows for 634 miles and drains the widest area (45,000 square miles). It is an extremely irregular river, with an outflow eight times greater in December and January than in August and September. Rising in the Massif Central on Mount Gerbier-de-Jonc, it flows northward over impervious terrain, with many gorgelike sections. Near Nevers it is joined by the Allier, another river of the massif. Within the Paris Basin the Loire continues to flow northward, as if to join the Seine system, but then takes a wide bend to the west to enter the Atlantic past Nantes and Saint-Nazaire. The Loire is artificially joined to the Seine by several

The Mediterranean-washed pebble beach at Nice on the French Riviera.

© Nedra Westwater/Blackstar

canals. The river's torrential flow, a hindrance to navigation, covers its floodplain with sand and gravel, which has commercial importance. The river is also a source of cooling water for a chain of atomic power stations near its course, which has raised concerns among environmentalists, as have various dam projects along the river. UNESCO designated the valley between Sully-sur-Loire and Chalonnes a World Heritage site in 2000.

The Garonne system. The Garonne, in the southwest, flows through the centre of the Aquitaine Basin. It is the shortest of the main French rivers, with a length of 357 miles, and it drains only 21,600 square miles. Its outflow is irregular, with high waters in winter (due to the oceanic rainfall) and in spring, when the snow melts, but with meagre flows in summer and autumn. Its source is in the central Pyrenees in the Aran (Joyeuse) Valley in Spain, and its main tributaries, the Tarn, the Aveyron, the Lot, and the Dordogne, originate in the Massif Central. With some exceptions, the whole network is generally useless for navigation and is filled with powerful, rapid, and dangerous currents.

Source of the Garonne

The Rhône system. In eastern France the direction of the main rivers is predominantly north-south through the Alpine furrow. The Rhône is the great river of the southeast. Rising in the Alps, it passes through Lake Geneva (Lac Léman) to enter France, which has 324 miles of its total length of 505 miles. At Lyon it receives its major tributary, the Saône. The regime of the Rhône is complex. Near Lyon the Rhône and its important Isère and Drôme tributaries, draining from the Alps, have a marked late spring–early summer peak caused by the melting of snow and ice. The course of the river and the local water tables have been much modified by a series of dams to generate power and to permit navigation to Lyon. The Rhône also supplies cooling water to a series of atomic power stations. West of the Rhône the Bas Rhône–Languedoc canal, constructed after World War II to provide irrigation, has proved to be an essential element in the remarkable urban and industrial development of Languedoc. East of the Rhône the Canal de Provence taps the unpolluted waters of a Rhône tributary, the Durance, supplying Aix-en-Provence, Marseille, Toulon, and the coast of Provence with drinking water and providing impetus for urban expansion. At its delta, beginning about 25 miles from the Mediterranean, the Rhône and its channels deposit significant amounts of alluvium to form the Camargue region.

The Rhine system. The Rhine forms the eastern boundary of France for some 118 miles. In this section its course is dominated by the melting of snow and ice from Alpine headstreams, giving it a pronounced late spring–summer peak and often generally low water in autumn. The Ill, which joins the Rhine at Strasbourg, drains southern Alsace. The Rhine valley has been considerably modified by the construction on the French side of the lateral Grand Canal d'Alsace, for power generation and navigation. The eastern Paris Basin is drained by two tributaries, the Moselle, partly canalized, and the Meuse.

The lakes. The French hydrographic system also includes a number of natural lakes of different origin. There are the lakes in depressions carved out by glaciation at the western periphery of the Alps, such as the lakes of Annecy and Bourget, the latter being the largest natural lake entirely within France. Others occur on the surfaces of ancient massifs and include the lakes of the Vosges. Some lakes are caused by structural faults and are lodged in narrow valleys, as are the Jura lakes. There are also lakes of volcanic origin, such as those in the Massif Central (crater lakes and lakes ponded behind lava flows), and regions scattered with lagoons or ponds, either created by coastal phenomena, as on the Landes (Atlantic) and Languedoc (Mediterranean) coasts, or caused by impervious terrain and poor local drainage, as in the Sologne plain. Major artificial lakes include the Serre-Ponçon reservoir, on the Durance River in the Alps, and the Sarrans and Bort-les-Orgues reservoirs, both in the Massif Central.

Alpine lakes

SOILS

On a broad, general scale, virtually the whole of France can be classified in the zone of brown forest soils, or brown earths. These soils, which develop under deciduous forest cover in temperate climatic conditions, are of excellent agricultural value. Some climate-related variation can be detected within the French brown earth group; in the high-rainfall and somewhat cool conditions of northwestern France, carbonates and other minerals tend to be leached downward, producing a degraded brown earth soil of higher acidity and lesser fertility; locally this may approach the nature of the north European podzol. The brown earth zone gives way southward to the zone of Mediterranean soils, which in France cover only a limited area. They are developed from decalcified clays with a coarse sand admixture and are typically red in colour because of the up-

© Rob Palmer from TSW—CLICK/Chicago

The château of Villandry, built in 1532, and its formal gardens in the Loire valley just east of Tours.

ward migration of iron oxides during the warm, dry summers. These soils can be quite fertile.

The *limon*

Over large areas of France, soils have developed not directly from the disintegrated bedrock but from the waste sheets created by periglacial action. These may provide a particularly favourable soil material; most notable is the windblown *limon* that mantles the Tertiary limestone plateaus of the central Paris Basin and the chalk beds to the northwest, the basis of the finest arable soils of France. Limestone and chalk enrich soils with lime, which is generally favourable, but there is a marked north-south contrast. The limestone areas of southern France tend to be swept almost bare of soil by erosion; the soil then collects in valleys and hollows. The soils of the higher mountains are naturally stony and unfavourable.

CLIMATE

Three major climatic zones

The climate of France is generally favourable to cultivation. Most of France lies in the southern part of the temperate zone, although the subtropical zone encompasses its southern fringe. All of France is considered to be under the effect of oceanic influences, moderated by the North Atlantic Drift to the west and the Mediterranean Sea to the south. Average annual temperatures decline to the north, with Nice on the Côte d'Azure at 59° F (15° C) and Lille on the northern border at 50° F (10° C). Rainfall is brought mainly by westerly winds from the Atlantic and is characterized by cyclonic depressions. Annual precipitation is more than 50 inches (1,270 millimetres) at higher elevations in western and northwestern France, in the western Pyrenees, in the Massif Central, and in the Alps and the Jura. In winter eastern France especially may come under the influence of the continental high-pressure system, which brings extremely cold conditions and temperature inversions over the cities, during which cold air is trapped below warmer air, with consequent fogs and urban pollution. The climate of France, then, can be discussed according to three major climatic zones—oceanic, continental, and Mediterranean, with some variation in the Aquitaine Basin and in the mountains.

PLANT AND ANIMAL LIFE

Plant life. Vegetation is closely related to climate, so that in France it is not surprising that there are two major but unequal divisions: the Holarctic province and the smaller Mediterranean province. Most of France lies within the Holarctic biogeographic vegetational region, characterized by northern species, and it can be divided into three parts. A large area of western France makes up one part. It lies north of the Charente River and includes most of the Paris Basin. There the natural vegetation is characterized by oak (now largely cleared for cultivation), chestnut, pine, and beech in uplands that receive more than 23.6 inches of annual rainfall. Heathland is also common, as a predominantly man-made feature (created by forest clearance, burning, and grazing). Broom, gorse, heather, and bracken are found. South of the Charente, the Aquitaine Basin has a mixture of heath and gorse on the plateaus and several varieties of oak, cypress, poplar, and willow in the valleys. On the *causses* of the Massif Central and on other limestone plateaus, broom, heath, lavender, and juniper appear among the bare rocks. The vegetation of eastern France, constituting a second part of the Holarctic division, is of a more central European type, with trees such as Norway maple, beech, pedunculate oak, and larch; hornbeam is often present as a shrub layer under oak. The various high mountain zones form a third Holarctic part; with cloudy and wet conditions, they have beech woods at lower elevations, giving way upward to fir, mountain pine, and larch but with much planted spruce. Above the tree line are high mountain pastures, now increasingly abandoned, with only stunted trees but resplendent with flowers in spring and early summer.

Vegetation of the Mediterranean zone

The second major vegetation division of the country lies within the Mediterranean climatic zone and provides a sharp contrast with the plant life elsewhere in France. The pronounced summer drought of this zone causes bulbous plants to die off in summer and encourages xerophytic plants that retard water loss by means of spiny, woolly, or glossy leaves; these include the evergreen oak, the cork oak, and all the heathers, cistuses, and lavenders. Umbrella, or stone, pine and introduced cypress dominate the landscape. The predominant plant life of the plateaus of Roussillon is the maquis, comprising dense thickets of drought-resistant shrubs, characterized in spring by the colourful flowers of the cistuses, broom, and tree heather; in most areas this is a form that has developed after human destruction of the evergreen forest. A large part of Provence's hottest and driest terrain is covered by a rock heath known as garigue. This region is a principal domain of the vineyard, but lemon and orange trees grow there also. At elevations of about 2,600 feet, as in the Cévennes, deciduous forest appears, mainly in the form of the sweet chestnut. At elevations of 4,500 feet this gives way to a subalpine coniferous forest of fir and pine.

Forest covers 58,000 square miles of France (15,000,000 hectares), which is more than a quarter of its territory. Most forests are on the upland massifs of the Ardennes and Vosges and within the Jura, Alps, and Pyrenees mountain chains, but extensive lowland forests grow on areas of poor soil, such as that of the Sologne plain south of the Loire River. The planted forest of maritime pine covering about 3,680 square miles (953,000 hectares) in the Landes of southwestern France is said to be the most extensive in western Europe. Increasingly, forests are less a source of wood and more a recreational amenity, especially those on the fringe of large urban agglomerations, such as Fontainebleau and others of the Île-de-France region.

Animal life. The fauna of France is relatively typical of western European countries. Among the larger mammals are red deer, roe deer, and wild boar, which are still hunted; the fallow deer is rather rare. In the high Alps are the rare chamoix and the reintroduced ibex. Hares, rabbits, and various types of rodents are found both in the forests and in the fields. Carnivores include the fox, the genet, and the rare wildcat. Among endangered species are the badger, the otter, the beaver, the tortoise, the marmot of the Alps, and the brown bear and the lynx of the Pyrenees. Seals have almost entirely disappeared from the French coasts. While French bird life is in general similar to that of its neighbours, southern France is at the northern edge of the range of African migrants, and such birds as the flamingo, the Egyptian vulture, the black-winged stilt, the bee-eater, and the roller have habitats in southern France.

People

ETHNIC GROUPS

The Basques

The French are, paradoxically, strongly conscious of belonging to a single nation, but they hardly constitute a unified ethnic group by any scientific gauge. Before the official discovery of the Americas at the end of the 15th century, France, located on the western extremity of the Old World, was regarded for centuries by Europeans as being near the edge of the known world. Generations of different migrants traveling by way of the Mediterranean from the Middle East and Africa and through Europe from Central Asia and the Nordic lands settled permanently in France, forming a variegated grouping, almost like a series of geologic strata, since they were unable to migrate any farther. Perhaps the oldest reflection of these migrations is furnished by the Basque people, who live in an isolated area west of the Pyrenees in both Spain and France, who speak a language unrelated to other European languages, and whose origin remains unclear. The Celtic tribes, known to the Romans as Gauls, spread from central Europe in the period 500 BC–AD 500 to provide France with a major component of its population, especially in the centre and west. At the fall of the Roman Empire, there was a powerful penetration of Germanic (Teutonic) peoples, especially in northern and eastern France. The incursion of the Norsemen (Vikings) brought further Germanic influence. In addition to these many migrations, France was, over the centuries, the field of numerous battles and of prolonged occupations before becoming, in the 19th and especially in the 20th century, the prime recipient of foreign immigration into Europe, adding still other mixtures to the ethnic melting pot.

LANGUAGES

French is the national language, spoken and taught everywhere. Brogues and dialects are widespread in rural areas, however, and many people tend to conserve their regional linguistic customs either through tradition or through a voluntary and deliberate return to a specific regional dialect. This tendency is strongest in the frontier areas of France. In the eastern and northern part of the country, Alsatian and Flemish (Netherlandic) are related to the Germanic languages; in the south, Occitan (Provençal or Languedoc), Corsican, and Catalan show the influence of Latin. Breton is a Celtic language related to languages spoken in some western parts of the British Isles (notably Wales), and Basque is a language isolate. Following the introduction of universal primary education during the Third Republic in 1872, the use of regional languages was rigorously repressed in the interest of national unity, and pupils using them were punished. More recently, in reaction to the rise in regional sentiment, these languages have been introduced in a number of schools and universities, primarily because some of them, such as Occitan, Basque, and Breton, have maintained a literary tradition. Recent immigration has introduced various non-European languages, notably Arabic.

RELIGION

About three-fourths of the French people belong to the Roman Catholic church. Only a minority, however, regularly participate in religious worship; practice is greatest among the middle class. The northwest (Brittany-Vendée), the east (Lorraine, Vosges, Alsace, Jura, Lyonnais, and the northern Alps), the north (Flanders), the Basque Country, and the region south of the Massif Central have a higher percentage of practicing Roman Catholics than the rest of the country. Recruitment of priests has become more difficult, even though the church, historically autonomous, is very progressive and ecumenical.

Protestant sects

Reflecting the presence of immigrants from North Africa, Algeria, and Morocco, France has one of Europe's largest Muslim populations: more than 4,000,000 Muslims, a sizable percentage of them living in and around Marseille in southeastern France, as well as in Paris and Lyon. Protestants, who number 700,000, belong to several different denominations. They are numerous in Alsace, in the northern Jura, in the southeastern Massif Central, and in the central Atlantic region. There are more than 700,000 adherents of Judaism, concentrated in greater Paris, Marseille, and Alsace and the large eastern towns. In addition to the religious groups, there also are several societies of freethinkers, of which the most famous is the French Masonry. Large numbers, however, especially among the working class and young population, profess no religious belief.

SETTLEMENT PATTERNS

Rural landscape and settlement. Centuries of human adaptation of the various environments of France have produced varied patterns of rural landscape. Scholars have traditionally made an initial contrast between areas of enclosed land (*bocage*), usually associated with zones of high rainfall and heavy soils, and areas of open-field land (*campagne*), generally associated with level and well-drained plains and plateaus. Two other patterns have evolved in the Mediterranean region and in the mountains.

Bocage. In its classic form, *bocage* is found in Brittany, where small fields are surrounded by drainage ditches and high earthen banks, from which grow impenetrable hedges arching over narrow sunken lanes. Similarly enclosed land is found elsewhere, however, notably on the northern, western, and southern fringes of the Paris Basin, such as in Normandy, as well as in the western and northern parts of the Massif Central, parts of Aquitaine, and the Pyrenean region. At higher levels hedges may be replaced by stone walls. Settlement mostly takes the form of hamlets and isolated farms.

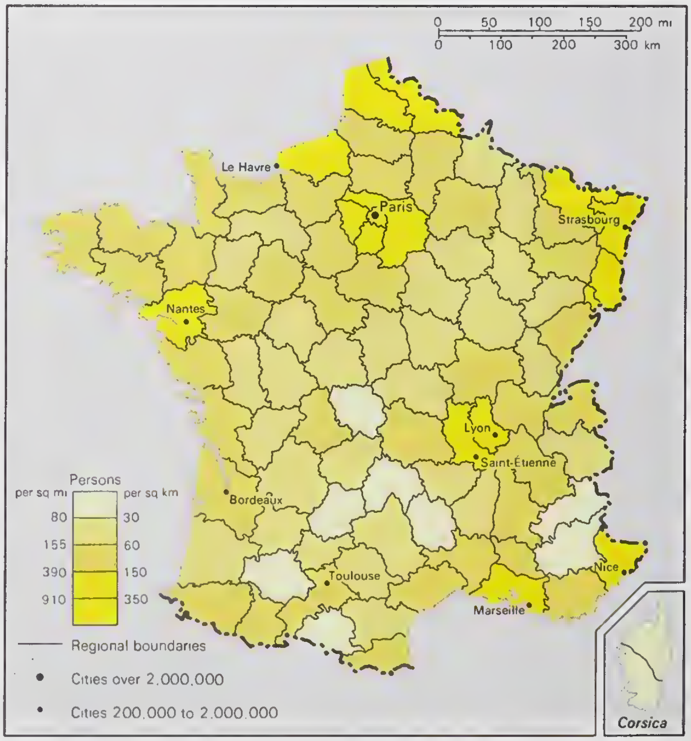
Open-field. The greatest extent of open-field land is found in the Paris Basin and in northern and eastern France, but there are pockets of it elsewhere. The landscape typically lacks hedges or fences; instead, the bewildering pattern of small strips and blocks of land is defined by small boundary stones. The land of one farmer may be dispersed in parcels scattered over a wide area. The land is predominantly arable, and the farmsteads are traditionally grouped into villages, which may be irregularly clustered or, as in Lorraine, linear in form.

Mediterranean. The generally block-shaped Mediterranean lowland parcels normally are not enclosed or are enclosed only by rough stone banks. However, in areas where delicate crops would be exposed to wind damage, there are screens of willows and tall reeds. Hillsides are frequently terraced, although much of this land type has been abandoned except in areas of intensive cultivation, such as the flower-growing region around Grasse. A very large farmhouse built on three floors is characteristic of wine-growing and sheep-raising regions, such as Provence. Rural population was formerly often clustered at high elevations, both for defense and in order to be above the malarial plains. In modern times there has been a move to more convenient lowland locations.

The *adrets* and *ubacs*

Mountain. In the high mountains and especially in the Alps, there is a contrast between the *adrets*, the sunny and cultivated valley slopes, and the *ubacs*, the cold and humid slopes covered with forests. The variety of vegetation on the slopes of the mountains is remarkable. Cultivated fields and grasslands are found in the depths of the valleys, followed in ascending order by orchards on the first sunny embankments, then forests, Alpine pastures, bare rocks, and, finally, permanent snow. A unique aspect of the mountain environment is that Alpine villages of the lower valley sides were often combined with *chalets* (*burons* in the Massif Central), temporary dwellings used by those tending flocks on summer pastures above the tree line.

Postwar transformation. After World War II the French government instituted a program of consolidation, whereby the scattered parcels of individual farmers were grouped into larger blocks that would accommodate heavier, mechanized cultivation. Initially progress was greatest in the open-field areas, particularly the Paris Basin, where there were few physical obstacles to the process. Subsequent extension to *bocage* areas had more severe consequences for landscape values and ecology, as hedges, sunken lanes, and ponds disappeared in favour of a new open landscape. At the same time, the vast numbers of people abandoning agricultural pursuits enormously changed the nature of



Population density of France.

rural settlement. Particularly in the more attractive areas, abandoned farms were purchased as second homes or for retirement. Where alternative employment was available, rural people stayed and became commuters, transforming barns and stables for other uses, such as garages. On the fringes of the expanding city regions, new houses and housing subdivisions for urban commuters were built in the villages, markedly changing their character.

Urban settlement. The primacy of Paris as the predominant urban centre of France is well known. After World War II the French government had an ambivalent attitude toward the development of the urban structure. On the one hand there was the desire to see Paris emerge as the effective capital of Europe, and on the other there was the policy of creating "*métropoles d'equilibre*," through which cities such as Lille, Bordeaux, and Marseille would become growth poles of regional development. Even more evident was the unplanned urbanization of small and medium-size towns related to spontaneous industrial decentralization from Paris, such as that along the Loire valley, or to retirement migration, such as that along the coastlands of southern France.

DEMOGRAPHIC TRENDS

In the second half of the 20th century, France's high postwar birth rate slowed, and about 1974 it fell into a sharp decline, eventually reaching a point insufficient for the long-term maintenance of the population. Since mid-century, because of a corresponding decline in the death rate, the rate of natural increase (balance of births against deaths) has remained positive, though declining. By the late 1990s France had an average population increase of more than 250,000 people each year. These changes were not exceptional to France; the same postwar pattern was largely paralleled in neighbouring countries. A number of factors combined to reduce the birth rate, among them the introduction of the contraceptive pill and the new preference for smaller families.

Emigration. Unlike many of its neighbours, France has never been a major source of international migrants, though in centuries past some left because of religious persecution (the Huguenots), while others settled in parts of France's colonial domain. Now, small numbers of French, especially from Brittany and Normandy, continue to relocate to Canada, and a number of Basques go to Argentina.

Immigration. From the middle of the 19th century there has been a substantial flow of immigrants into France, with waves of immigrants arriving especially after the two World Wars. France had the reputation into the early 20th century of being the European country most open to immigrants, including political refugees, but this reputation changed in the late 20th century, when opposition rose to continued immigration from Africa. Although immigration flattened out after 1974, natural increase dropped, so that immigration continued to contribute significantly to population growth. At the end of the century, there were more than three million foreigners resident in France, amounting to some 6 percent of the population, a proportion that had remained constant since 1975. Neighbouring countries such as Portugal, Italy, and Spain continued to be significant contributors, but recent immigrant streams came from North Africa, notably Algeria (an integral part of France until 1962) and the former protectorates of Morocco and Tunisia. Peoples from French or former French territories in Central Africa, Asia, and the Americas provided an additional source of immigrants.

Effects of wars on population trends

As the numbers of immigrants grew, so did incidents of racial discrimination in housing and employment, as well as social activism among immigrant groups. Initially, immigrants from Africa and the Americas were predominantly males, living in low-standard housing and working in undesirable, low-skilled occupations. As families were progressively reconstituted, immigrants continued to work in jobs that Frenchmen were reluctant to accept. With the beginning of an economic downturn in 1974, though, French workers began to reclaim some of the jobs held by immigrants, and the government began to restrict immigration. Adding to the job competition were approximately one million persons with French citizenship, the so-called *pieds-noirs* (literally "black feet"), who were repatriated from territories in North Africa decolonized in 1962–64. The policy of restricting immigration remains in force, with the result that in the early 21st century the net annual increase of population from immigration averaged little more than 50,000 people. With the enactment in 1999 of the Amsterdam Treaty in France, many issues of immigration became shared by participating members of the European Union.

Repatriation from North Africa

Population structure. The aging of the population is common to western Europe, but because of low birth rates it has been observable in France since the beginning of the 19th century. By the end of the 20th century, one-fifth of French citizens were at least 60 years old. The tendency for the proportion of the elderly population to increase also reflects medical advances, which have produced a longer expectation of life. The age structure of the population is of considerable social and economic importance. The steady increase in the proportion of the aged puts an increasing strain on the working population to provide pensions, medical and social services, and retirement housing. The increase in births between 1944 and the mid-1970s, however, brought its own problems, notably the need to rush through a school-building program, followed by the creation of new universities. But this demographically young population also stimulated the economy by creating a greater demand for consumer goods and housing.

Another important aspect of population structure is the proportion of men to women, in society as a whole and in the various age groups. As in most western European countries, women outnumber men in French society, particularly in the older age groups, which is the result of two factors: the wars, which caused the death of a large number of men, and the natural inequality of life expectancy for men and women. A French woman at birth has one of the highest life expectancies in the world (83 years), while a man's is much lower (75 years), although still relatively high when compared with the world in general. The ratio of men to women in employment is another measure of population structure, and in the late 20th century women steadily increased their share of the job market.

Proportion of men to women

Population distribution. Particularly low population densities are characteristic of the mountain regions, such as the Massif Central, the southern Alps, the Pyrenees, and Corsica, but are also reflected in some lowland rural areas, such as the eastern and southern Paris Basin and large parts of Aquitaine. The four least-populated French *régions*—Corsica, Limousin, Franche-Comté, and Auvergne—have one-sixteenth of the national population in about one-eighth of the area. By contrast, the four most populated French *régions*—Île-de-France (Paris region), Rhône-Alpes, Provence-Alpes-Côte d'Azur, and Nord-Pas-de-Calais—have more than two-fifths of the French population in less than one-fifth of the area. Other high-density areas are the industrial cities of Lorraine; isolated large cities, such as Toulouse; and certain small-farm areas, such as coastal Brittany, Flanders, Alsace, and the Limagne basin of Auvergne.

Until about the mid-19th century, rural and urban populations both increased; thereafter there was a marked depopulation of the more remote, mostly mountainous, rural areas and a swing to urban growth. In the space of a century, from the 1860s to the 1960s, rural population decreased by more than one-third, though since that time it has remained constant. There were still as many rural as urban inhabitants even up to the period between the two World Wars, but since the 1980s three-fourths of the population has been urban. Postwar rural depopulation was associated with the exodus of labour following the modernization of French agriculture and the growth of urban industrial regions.

Rural depopulation

From about 1975, the older industrial *régions*, such as Nord-Pas-de-Calais and Lorraine, were in decline and became regions of out-migration. The most dramatic reversal was that of the deindustrialized Île-de-France *région*; although still the greatest population concentration of France, it had a negative migration balance after 1975. Growth subsequently switched to the south, to the coastlands of Languedoc and of Provence-Alpes-Côte d'Azur;

to the west, in the Atlantic *régions* of Poitou-Charentes and Pays de la Loire; and to the southwest, in the Midi-Pyrénées and Aquitaine *régions*. These shifts reflect a combination of economic decentralization, retirement migration, sunbelt industrialization, changing residential preferences, and expanding tourism. Population increase has also been strong on the southern and western fringes of the Paris Basin, favoured for industrial decentralization from the Île-de-France *région*.

Suburbanization

In the first half of the 20th century, suburban growth, where it did occur, was not the result of middle-class suburbanization, as it was in the United States and the United Kingdom. It was the working class and the lower middle class that moved out of Paris, while higher-income groups endeavoured to maintain a foothold in the central city. In the postwar period, however, suburbanization took increasingly middle-class forms, with the building up of satellite low-density subdivisions known as "new villages." Similar postwar suburbanization occurred in cities such as Marseille, Lyon, Lille, and Bordeaux.

Increasingly, the most rapid population growth is relegated to small towns and nominally rural communes on the expanding fringes of the city regions. This dispersal of population is associated with an increasing length of daily commuter movements, with all their human disadvantages, as well as other problems of urban sprawl.

(T.H.El./J.N.T./Ed.)

Economy

France is one of the major economic powers of the world, ranking along with such countries as the United States, Japan, Germany, Italy, and the United Kingdom. Its financial position reflects an extended period of unprecedented growth that lasted for much of the postwar period until the mid-1970s; frequently this period was referred to as the *trente glorieuses* ("thirty years of glory"). Between 1960 and 1973 alone, the increase in gross domestic product (GDP) averaged nearly 6 percent each year. In the aftermath of the oil crises of the 1970s, growth rates were moderated considerably and unemployment rose substantially. By the end of the 1980s, however, strong expansion was again evident. This trend continued, although at a more modest rate, through the end of the century.

Despite the dominance of the private sector, the tradition of a mixed economy in France is well established. Successive governments have intervened to protect or promote different types of economic activity, as has been clearly reflected in the country's national plans and nationalized industries. In the decades following World War II, the French economy was guided by a succession of national plans, each covering a span of approximately four to five years and designed to indicate rather than impose growth targets and development strategies.

The public sector in France first assumed importance in the post-World War II transition period of 1944–46 with a series of nationalizations that included major banks such as the National Bank of Paris (Banque Nationale de Paris; BNP) and Crédit Lyonnais, large industrial companies such as Renault, and public services such as gas and electricity. Little change took place after that until 1982, when the then-Socialist government introduced an extensive program of nationalization. As a result, the enlarged public sector contained more than one-fifth of industrial employment, and more than four-fifths of credit facilities were controlled by state-owned banking or financial institutions. Since that period successive right-wing and, more recently, left-of-centre governments have returned most enterprises to the private sector; state ownership is primarily concentrated in transport, defense, and broadcasting.

Improved living standards

Postwar economic growth has been accompanied by a substantial rise in living standards. As incomes have risen, proportionately less has been spent on food and clothing and more on items such as housing, transportation, health, and leisure. Workers' incomes are taxed at a high to moderate rate, and indirect taxation in the form of a value-added tax (VAT) is relatively high. Overall, taxes and social security contributions levied on employers and employees in France are higher than in many other European countries.

AGRICULTURE, FORESTRY, AND FISHING

France's extensive land area—of which more than half is arable or pastoral land and another quarter is wooded—presents broad opportunities for agriculture and forestry. The country's varied relief and soils and contrasting climatic zones further enhance this potential. Rainfall is plentiful throughout most of France, so water supply is not generally a problem. An ample fish supply in the Atlantic Ocean and Mediterranean Sea provides an additional resource.

Agriculture employs relatively few people—about 4 percent of the labour force—and makes only a small contribution to GDP—about 3 percent. Yet France is the EU's leading agricultural nation, accounting for more than one-fifth of the total value of output, and alone is responsible for more than one-third of the EU's production of oilseeds, cereals, and wine. France also is a major world exporter of agricultural commodities, and approximately one-eighth of the total value of the country's visible exports is related to agriculture and associated food and drink products.

Agricultural land use

France has a usable agricultural area of nearly 74 million acres (30 million hectares), more than three-fifths of which is used for arable farming (requiring plowing or tillage), followed by permanent grassland (about one-third) and permanent crops such as vines and orchards (about one-twentieth). Areas in which arable farming is dominant lie mostly in the northern and western regions of the country, centred on the Paris Basin. Permanent grassland is common in upland and mountainous areas such as the Massif Central, the Alps, and the Vosges, although it is also a notable feature of the western *région* of Basse-Normandie. Conversely, the major areas devoted to permanent cultivation lie in Mediterranean regions.

Grains. More than half of the country's arable land is used for cereals, which together provide about one-sixth of the total value of agricultural output. Wheat and corn (maize) are the main grains, with other cereals, such as barley and oats, becoming progressively less important. There are few areas of the country where cereals are not grown, although the bulk of production originates in the Paris Basin and southwestern France, where natural conditions and (in the former case) proximity to markets favour such activity. A considerable area (about one-seventh of the agricultural area), predominantly in western France, is also

© Serraillier—Rapho/Photo Researchers

Harvesting grapes in a vineyard at Ay, near Épernay, in the Champagne region.

given over to forage crops, although the acreage has been shrinking since the early 1980s as dairy herds have been reduced in accordance with EU guidelines. In contrast, there has been a substantial increase in oilseed output; the area under cultivation has quadrupled since the early 1980s and now approaches one-tenth of agricultural land.

Fruits and wine making. Vines, fruits, and vegetables cover only a limited area but represent more than one-fourth of the total value of agricultural output. France is probably more famous for its wines than any other country in the world. Viticulture and wine making are concentrated principally in Languedoc-Roussillon and in the Bordeaux area, but production also occurs in Provence, Alsace, the Rhône and Loire valleys, Poitou-Charentes, and the Champagne region. There has been a marked fall in the production of vin ordinaire, a trend related to EU policy, which favours an increase in the output of quality wines. Fruit production (mainly of apples, pears, and peaches) is largely concentrated in the Rhône and Garonne valleys and in the Mediterranean region. Vegetables are also grown in the lower Rhône and Mediterranean areas, but a large part of output comes from western France (Brittany) and the southwest and the northern *régions* of Nord–Pas-de-Calais and Picardy, where sugar beets and potatoes are produced.

Dairying and livestock. Cattle raising occurs in most areas of the country (except in Mediterranean regions), especially in the more humid regions of western France. Animal-related production accounts for more than one-third of the total value of agricultural output. In general, herds remain small, although concentration into larger units is increasing. Overall, however, the number of cattle has been falling since the early 1980s, largely as a result of EU milk quotas. These have adversely affected major production areas such as Auvergne, Brittany, Basse-Normandie, Pays de la Loire, Rhône-Alpes, Lorraine, Nord–Pas-de-Calais, and Franche-Comté. One result has been an increasing orientation toward beef rather than dairy breeds, notably in the area of the Massif Central. The raising of pigs and poultry, frequently by intensive methods, makes up more than one-tenth of the value of agricultural output. Production is concentrated in the *régions* of Brittany and Pays de la Loire, encouraged originally by the availability of by-products from the dairy industry for use as feed. Sheep raising is less important. Flocks graze principally in southern France on the western and southern fringes of the Massif Central, in the western Pyrenees, and in the southern Alps.

Reduction in cattle numbers

Agribusiness. Agriculture has changed in other ways. Farm structures have been modified substantially, and the number of holdings have been greatly reduced since 1955, numerous small farms disappearing. By the late 1990s there were fewer than 700,000 holdings, compared with more than 2,000,000 in the mid-1950s. The average size of farms has risen considerably, to nearly 100 acres (40 hectares). Large holdings are located primarily in the cereal-producing regions of the Paris Basin, while small holdings are most common in Mediterranean regions, the lower Rhône valley, Alsace, and Brittany. Important technical changes have also occurred, ranging from the increased use of intermediate products such as fertilizers and pesticides to the widespread use of irrigation (nearly one-tenth of agricultural land is now irrigated) and the growth of crops within controlled environments, such as under glass or plastic canopies. Marketing systems have also been modified, as an increasing proportion of output is grown under contract. Together such changes have led to a remarkable increase in output of major agricultural products, but they have also resulted in a large reduction in the number of agricultural workers and the increased indebtedness of many farmers, and the related negative effects on the environment have given rise to the organic farming movement.

Forestry. With more than 57,000 square miles of woodland, France possesses one of the largest afforested areas in western Europe, offering direct employment to more than 80,000 people. Forested areas are unevenly distributed, with the majority lying to the east of a line from Bordeaux to the Luxembourg border. Aquitaine and Franche-Comté have a particularly dense forest cover. This vast resource is, however, generally underexploited, partly because of the multitude of private owners, many of whom are uninterested in the commercial management of their estates. Less than one-fourth of the afforested area is controlled by the National Office of Forests.

Fishing. Despite the extent of France's coastlines and its numerous ports, the French fishing industry remains relatively small. Annual catches have averaged about 700,000 tons since the mid-1970s, and by the end of the 1990s there were fewer than 16,500 fishermen. The industry's problems are related to its fragmented character and to inadequate modernization of boats and port facilities, as well as to overfishing and pollution. Activity is now concentrated in the port of Boulogne in Nord–Pas-de-Calais and to a lesser degree in ports in Brittany such as Concarneau, Lorient, and Le Guilvinic. France is also known for its aquaculture, with activity increasing over recent years along the coastal waters of western France. Oyster beds are found particularly in the southwest, centred on Marennes-Oléron.

Major fishing ports

RESOURCES AND POWER

Compared with its agricultural resources, the country is far less well endowed with energy resources. Coal reserves are estimated at about 140 million tons, but French coal suffered from being difficult and expensive to mine and from its mediocre quality. In 1958 annual production amounted to some 60 million tons; 40 years later this total had dropped to less than 6 million tons, and in 2004 the last coal mine was shuttered. Consequently, imported coal long supplemented indigenous production and greatly exceeded domestic output. Imports originate mainly from Australia, the United States, South Africa, and Germany.

Other energy resources are in short supply. Natural gas was first exploited in southwestern France (near Lacq) in 1957. Production then increased substantially, only to decline after 1978 as reserves became exhausted. By the late 1990s, production was negligible, requiring a high level of imports, principally from the North Sea (Norway and The Netherlands), Algeria, and Russia. France has few oil reserves, and production from wells in Aquitaine and the Paris Basin is extremely limited. Uranium is mined in the Massif Central, and, although recoverable reserves are estimated at approximately 50,000 tons, more than half of the annual consumption has to be imported. France, however, does possess fast-moving rivers flowing out of highland areas that provide it with an ample hydroelectric resource.

Minerals. The metal industry is poorly supplied by indigenous raw materials, although traditionally France was an important producer of iron ore and bauxite. Iron ore output exceeded 60 million tons in the early 1960s, originating principally in Lorraine; but production has now ceased, despite the continued existence of reserves. Low in metal content and difficult to agglomerate, Lorraine ores were thus long supplemented and have now been replaced by richer overseas supplies from such countries as Brazil, Sweden, and Australia. Bauxite production is negligible, though other mineralized ores, such as those containing lead, zinc, and silver, are mined in very small quantities. Greater amounts of potash (mined in Alsace), sodium chloride (from mines in Lorraine and Franche-Comté and from salt marshes in western and southern France), and sulfur (derived from natural gas in Aquitaine) are produced, but again the trend is toward declining output as reserves are depleted. The supply of stone, sand, and gravel is relatively ubiquitous.

Energy. Through the post-World War II years, the increase in the demand for energy has closely followed the rate of economic growth. Thus, for much of the period until 1973, consumption increased rapidly. Then, in the wake of the two oil price rises of 1973 and 1979, demand stabilized, followed by a fall in the early 1980s until growth rates recovered after the mid-1980s.

The demand for different types of energy has changed considerably over time. In the early postwar years, coal provided the larger part of energy needs. By the 1960s, however, oil, as its price fell in real terms, was being used in ever-greater quantities, so that by 1973 about two-thirds of energy consumption was accounted for by crude oil. Since then a more diversified pattern of use has emerged. Coal now plays only a minor role, while the use of oil has also fallen, replaced partly by natural gas and notably by nuclear energy, which now accounts for more than one-third of primary energy consumption. One of the main consequences of these changes has been a reduction in the country's previously high dependence on external sources of supply.

Oil has long been France's principal energy import, which has led to the growth of a major refining industry, with plants concentrated in two areas of the lower Seine valley (Le Havre and Rouen) and in the region around Fos-sur-Mer and the Étang de Berre. Many markets are supplied with oil products by pipeline, which is also the distribution method for natural gas. Algerian imports arrive in the form of liquefied natural gas (primarily methane) and are unloaded at French ports where regasification plants operate.

Nuclear energy

Since the early 1980s one of the most significant changes in energy supply has been the greatly increased role of nuclear power, at the expense of fuel oil and coal; even the production of hydroelectric power has stabilized, as most suitable sites have already been exploited, particularly those of the Rhine and Rhône valleys, the Massif Central, and the Alps. In contrast, nuclear production, benefiting from major government investment from the early 1970s, expanded enormously in the 1980s, notably with the construction of sites in the Rhône and Loire valleys, a reflection of the need for large quantities of cooling water. By the late 1990s more than three-fourths of electricity in France was produced in nuclear plants, the highest proportion in the world, which enabled the country to become a large exporter of such energy. More recently development has slowed substantially, as demand has eased and environmental groups have campaigned against further investment. France's nuclear industry also includes a large uranium-enrichment factory at Pierrelatte in the lower Rhône valley and a waste-reprocessing plant at La Hague, near Cherbourg.

MANUFACTURING

Industrial trends. French industry was long the powerhouse of the country's postwar economic recovery. Yet, after a period of substantial restructuring and adjustment, particularly during successive periods of recession since the late 1970s, this sector (including construction and civil engineering) now employs only about one-fourth of the country's workforce and contributes the same proportion of GDP.

Changes in industrial location have also occurred. Industrial expansion in the 1960s and '70s was accompanied by large-scale decentralization, favouring many areas of the Paris Basin (where there was an abundant and relatively cheap supply of labour) at the expense of the capital. Few company headquarters followed the dispersion of manufacturing plants, however, so that the centre of industrial operations remained rooted in the Paris region. The decline of industrial employment since the mid-1970s has had the greatest impact in traditional manufacturing regions, such as Nord–Pas-de-Calais and Lorraine. Nevertheless, the broad arc of *régions* stretching through northern and eastern France, from Haute-Normandie to Rhône-Alpes, remains the most heavily industrialized part of the country.

The automobile industry

Branches of manufacturing. On the basis of employment and turnover, seven branches of manufacturing stand out as particularly important: vehicles, chemicals, metallurgy, mechanical engineering, electronics, food, and textiles. The vehicle industry is dominated by the activities of the two automobile manufacturers, Peugeot SA (including Citroën) and Renault, which together produced nearly four million cars annually at the beginning of the 21st century. Automobile production generates a substantial number of direct jobs as well as employment in subsidiary industries, such as the major tire manufacturer Michelin. France also possesses an important industry for the manufacture of railway locomotives and rolling stock, for which the expanding high-speed train (*train à grande vitesse*; TGV) network represents a major market.

Within the chemical industry, manufacturing ranges from basic organic and inorganic products to fine chemicals, pharmaceuticals, and other parachemical items, including perfumes. Because of the capital-intensive nature of these activities, a dominant role is played by large manufacturers such as Rhône-Poulenc. Extensive research is carried out in this field. Basic chemical production is concentrated in areas offering access to raw materials, such as Nord–Pas-de-Calais, Étang-de-Berre, and Rhône-Alpes, whereas pharmaceutical production is more closely related to major market areas and research centres, notably Île-de-France.

The metallurgical industry, dominated by the production of steel, experienced major restructuring in the late 1970s and the '80s as demand fell and competition from other international producers increased. Originally concentrated in Lorraine because of the presence of iron ore, steel production shifted to the coastal sites of Dunkirk and Fos-sur-Mer, which relied on imported ore and coal. France is also an important producer of aluminum, notably through the Pechiney group. Such basic metal industries support a diverse range of engineering activities, spread widely throughout France but with important concentrations in the highly urbanized and industrialized *régions* of Île-de-France and Rhône-Alpes. Similar features characterize the electrical engineering and electronics industries. France is a major manufacturer of professional electronics, such as radar equipment, but is weakly represented in the field of consumer electronics, which has led to a high level of imports. The country also has a number of high-tech aerospace industries, which manufacture aircraft, missiles, satellites, and related launch systems. These industries are concentrated in the Paris region and in the southwest around Toulouse and Bordeaux.

Food and beverage production

Food and beverage industries represent the largest branch of French manufacturing, reflecting the considerable volume and diversity of agricultural production. Although present in most regions, food manufacturers are particularly concentrated in major urban market areas and in western agricultural regions such as Brittany, Pays de la Loire, and Basse-Normandie. The beverage sector is dominant in the main wine-growing areas of northern and northeastern France; it represents an important source of exports.

Textile and clothing industries have experienced a long period of decline in the face of strong foreign competition, with substantial job losses and plant closures affecting the major production areas of northern France and Rhône-Alpes (textiles), as well as Île-de-France (clothing). Unlike other major industrial branches, these activities remain characterized by small firms.

A varied group of construction and civil engineering industries employs about one-fourth of the labour in the industrial sector. Activity and employment have fluctuated considerably in relation to changing government and private investment programs and the varying demand for new homes. This sector is characterized by the coexistence of a large number of small firms with a limited number of large companies, many of which work on civil engineering contracts outside France.

FINANCE

Growth of branch banking

Banking and insurance. France possesses one of the largest banking sectors in western Europe, and its four major institutions, Crédit Agricole, BNP-Paribas, Crédit Lyonnais, and Société Générale, rank among the top banks on the continent. Traditionally, banking activities were tightly controlled by the government through the Banque de France. However, deregulation beginning in the 1960s led to a substantial increase in branch banking and bank account holders, and legislation in 1984 further reduced controls over banks' activities, which thereby enabled them to offer a wider range of services and led to greater competition. Since then, encouraged by the lifting of restrictions on the free movement of capital within the EU in 1990, banks have broadly internationalized their activities. In 1993 the Banque de France was granted independent status, which freed it from state control. In general, employment in the banking sector has declined, largely because of the widespread computerization of transactions and this restructuring. At the turn of the 21st century, the franc gave way to the euro as the legal currency in France.

France has a large insurance industry dominated by major companies such as Axa, CNP, and AGF but also including a number of important mutual benefit societies, which administer pension plans. The deregulation of this sector has led to vast reorganization, with activity still concentrated in Paris though a number of provincial towns have developed as specialist centres through the location of various mutual societies.

The stock exchange. Share transactions in France are centred on the Bourse de Paris (Paris Stock Exchange), a national system that since 1991 has incorporated much smaller exchanges at Lyon, Bordeaux, Lille, Marseille, Nancy, and Nantes. Share dealings and stock market activity have increased greatly since the early 1980s, corresponding with a period of deregulation and modernization: official brokers lost their monopoly on conducting share transactions; a second market opened in 1983 to encourage the quotation of medium-size firms; and in 1996 the "new market" was launched to help finance young, dynamic companies in search of venture capital. Also in 1996 the Bourse was restructured, reinforcing the powers of its controlling body, Commission des Opérations de Bourse.

Foreign investment. Financial deregulation, the movement toward a single European market, and the general freeing of world trade are among the influences that have encouraged investment by French firms outside France and increased the reverse flow of foreign investment funds into the country. In the industrial field French companies have shown a growing interest in investing in other advanced economies, especially the United States. Over recent years investments have also multiplied in the developing economies of Asia and eastern Europe. Foreign firms investing in France have been principally from the EU (notably the United Kingdom, The Netherlands, and Germany) and the United States. Most investment is related to the fields of engineering, electronics, and chemicals and generally is directed at the more highly urbanized centres of the country. The sources and nature of foreign investment in France are becoming more diverse, however. Japanese interests have increased substantially, for instance, and investment in property and the service industry has been growing, particularly in and around Paris.

TRADE

France, a leading trading nation, has grown into one of the world's foremost exporting countries, with the value of exports representing more than one-fifth of GDP. France is also a major importer, especially of machinery, chemicals and chemical products, tropical agricultural products, and traditional industrial goods such as clothes and textiles. The high level of imports led to a trade deficit for much of the period between the early 1970s and early 1990s. However, from 1992 France experienced a trade surplus, combined with a positive balance from invisible (nonmerchandise) transactions, especially tourism.

Most foreign trade is based on the exchange of goods. In the case of agricultural commodities, France has become an increasingly important net exporter of raw agricultural products (such as grains) as well as agro-industrial products, such as foods and beverages, including wines, tinned fruits and vegetables, and dairy products. The need to import large quantities of oil (and to a lesser extent gas and coal), however, has resulted in a sizable deficit for those exchanges. Although France imports a great deal of industrial goods, the country has long been a major exporter of vehicles and transport equipment, as well as armaments and professional electronics. More recently exports of pharmaceuticals and parachemical products have risen.

Principal trading partners

The greater part of foreign trade is carried out with other developed countries, and some four-fifths of transactions take place with Organisation for Economic Co-operation and Development (OECD) countries. Among these the EU plays a major role, reflecting the growing exchange of goods and services between its member countries. More than three-fifths of French exports and imports are destined for or originate in EU countries, of which Germany is easily the most important. Outside the EU the United States is France's other major trading partner. EU countries are an important source of industrial imports, whereas fuel products and raw materials tend to originate from more distant sources. Conversely, agricultural and food exports are oriented predominantly toward European markets, whereas industrial goods are exported to a more global marketplace.

SERVICES

Tertiary-sector growth

The various service, or tertiary, industries in France account for about two-thirds of the country's employment and of GDP. These levels were reached following an extended period of sustained growth, notably since the 1960s. This sector covers a highly diverse range of activities, including social and administrative services, such as local government, health, and education; wholesaling, distribution, and transport and communication services; consumer services, such as retailing and the hotel and catering trades; and producer or business services, including banking, financial, legal, advertising, computing, and data-handling services.

Not all tertiary activities have developed in the same way. For example, rationalization in the banking and financial services sector has limited the creation of jobs. Conversely, the continuously strong growth, since the early 1970s, of hypermarkets and other large frcestanding retail outlcts that allow for purchasing in bulk and in greatcr variety has led to a significant rise in related employment. In particular the large group of producer services has expanded rapidly. In part this trend is the inevitablc consequence of the increasingly complex and highly competitivc nature of the modern economy. It also results from companies' strategies of externalizing (outsourcing) such service rcquircments for reasons of efficiency and cost savings.

Central-ization in Paris

Tertiary activities are located predominantly in urban areas, especially the larger citics. Such concentration is most evident in relation to thc capital. The Île-dc-France *région* surrounding Paris accounts for ncarly one-fourth of all tertiary employment alone, despite containing less than one-fifth of the population. In Paris the scctor's importance is qualitative as well as quantitative. Paris houscs more than two-thirds of the headquarters of the country's major companies and a disproportionatcly largc share of senior managcment and rcsearch staff. This attraction to the capital is influenced by a numbcr of factors, including the size and diversity of the labour market, thc high level of accessibility to other French and international business centres, prestige, and the prcsence of numerous spccialized services.

Civil service. The largest groups of employces are those in national education and thc postal systcm. As in the judicial system, French administration has bccn strongly marked by a strict hierarchy since the time of Napolcon. Civil servants are grouped into diffcrent corps and different ranks and are classified according to their rccruitmcnt level into four different categories. Entry is by a compctitive examination. At the highest level, category A civil servants are recruited through a national school of administration, created in 1945, which gives access to the *grands corps de l'État*, including the Court of Accounts, the Inspection of Finance, the prcfectural corps, the diplomatic service, and the civil administrators' corps. The duties and rights of civil servants are defined by a general statute of 1946, which was partly modified in 1959. The career guarantees and disciplinary code are extensive and are protected by the Conseil d'État (Council of State). In return, civil servants are duty-bound to be discreet in expressing any personal opinions, and the right to strike, which is recognized by the constitution for all French citizens, is severely limited for them, although this varics according to the corps. Most civil servants belong to labour unions.

Tourism. With France's variety of landscapes and climatic conditions, its cultural diversity, and its renowned cuisine, it is of little surprise that tourism should have become a major industry. Directly and indirectly this activity employs about 7 percent of the workforce and contributes approximately 8 percent of GDP, earning French businesses a substantial income from foreign visitors and more than compensating for the amount spent by French tourists abroad. France is one of the world's leading tourist destinations, visited by up to 70 million foreign tourists each year at the end of the 20th century.

The tourist industry has grown rapidly since the 1960s, with an increasingly large number of French families taking a holiday each year, encouraged by greater affluence, more leisure time, and, since 1982, five weeks' statutory paid holiday. In response to this increase in demand, the industry itself has changed. An activity traditionally distinguished by small businesses has been transformed by the growth of increasingly large hotel and holiday firms; new resorts have been built, notably along the Languedoc and Aquitaine coasts and in the French Alps, and new tourist products have been developed, including spectacular theme parks. The Disneyland complex on the eastern fringe of Paris, which opened in 1992, epitomized this trend.

Comparatively few French people take their holidays abroad. Conversely, France receives a large influx of foreign visitors, mainly from European countries, especially Germany. The unequal impact of tourism on different regions is a key feature of this activity. In summer a restricted number of coastal areas, notably in the Midi and in Brittany, receive the heaviest influx of holidaymakers; in winter mountainous regions become the preferred destination, particularly the northern Alps, with such major ski resorts as Chamonix, Tignes, La Plagne, and Les Arcs. Paris itself is an enormous tourist attraction, especially for foreign visitors and for events such as exhibitions and conferences; indeed, the capital is perhaps the world's leading centre for international conferences.

LABOUR AND TAXATION

Structural changes in the economy have helped transform the French labour force. Since the 1960s there has been a growing transfer from blue- to white-collar occupations, particularly as jobs in management, the professions, and administration have greatly increased. This change has been accompanied by a marked rise in female employment, so that almost half of all jobs are now held by women. A significant increase in part-time work and employment on fixed-term contracts has also taken place for both sexes. Firms have favoured this development because of the greater flexibility it offers, as have employees themselves, seeking freer, less-formalized working arrangements. The trend has also been encouraged by short-term government measures to reduce unemployment.

Such changes away from standard jobs have also contributed to the weakened position of trade unions in France: as little as one-tenth of French workers belong to a union. Traditional support from blue-collar workers has also been eroded by heavy job losses in industries such as steel, shipbuilding, and vehicles. The main trade unions are the General Confederation of Labour, Force Ouvrière (literally "workforce"), and French Democratic Confederation of Labour. With the exception of those in 1968, major nationwide strikes have been relatively infrequent in France. Employers, for their own part, are grouped together within the Movement for French Enterprises (Mouvement des Entreprises de France), which in 1998 replaced the National Council of French Employers (Conseil National du Patronat Français). This organization represents all firms in negotiations with the government, state administrative services, and unions.

TRANSPORTATION AND TELECOMMUNICATIONS

The transportation sector includes such dynamic companies as the National Society of French Railways (Société Nationale des Chemins de Fer Français), the state-owned railways operator, and Air France, the national airline. Closely allied are manufacturers of transport equipment and the civil engineering concerns responsible for constructing new infrastructure. Generally, France benefits from a dense and diversified transport network, limited only by its still excessive focus upon the capital city. For land-based movements the road network has become increasingly important. For example, a vast majority of all freight traffic, in terms of the volume and distance of goods moved, goes by road. This dominance has been achieved at the expense of railways and inland waterways.

Growth of highway traffic

Roads. Traffic on the highways has more than doubled since 1970, and about one-fifth of vehicles are commercial. An extensive road system totaling about 600,000 miles has been developed to deal with increasingly heavy traffic conditions. However, only a small proportion of this network consists of main trunk roads (the *routes nationales*) and motorways. Construction on the latter began much later than in neighbouring countries, and it was not until the mid-1960s that a major development program was under way. To speed progress, building concessions were granted to private and semiprivate companies, which, in return for their investment, were authorized to levy tolls. Since that period the major radial routes from the capital have been completed, as have embryonic regional networks focusing on large urban centres, such as Paris, Lyon, Marseille, and Lille. Traffic is heavily concentrated on the main north-south axis between these cities. In extending the system, emphasis has been placed on improving international links and developing national routes that avoid Paris, as between Calais and Dijon, as well as Bordeaux and Clermont-Ferrand. Numerous rural roads and lanes supplement the main system.

Railroads. By the end of the 19th century, the present rail network was largely in place, dominated by the main lines radiating from Paris. Since World War II many little-used rural sections have been closed. In contrast, since the early 1980s certain new lines have been opened in conjunction with the introduction of high-speed passenger trains (*trains à grande vitesse*; TGV) between Paris and a number of provincial cities. Southeastern France was the first area to be provided with such services, reflecting the already high density of traffic between Paris, Lyon, and the Mediterranean coast. New lines are also in operation to western and northern France, with longer-term plans to serve eastern regions. International service also exists to Geneva, Lausanne, and Brussels, as well as to London, by means of the Channel Tunnel, which opened in 1994 after six years of construction. It is used for passenger and freight trains as well as for transporting cars and commercial vehicles. By the end of the 20th century, the Eurostar passenger trains linked Paris to London in three hours and carried more than seven million travelers annually. In France the TGV network alone accounts for more than one-half of passenger miles and has attracted many new customers to the railways. Generally, however, fewer than one-fifth of passenger movements in France were accounted for by rail services, with traffic heavily concentrated along the main, electrified radial routes from the capital, particularly in the direction of southeastern France. Freight traffic has declined, partly because of fallen demand for products such as coal, iron, and oil, traditionally carried by rail, and partly because of intense competition from road haulers. Like passenger traffic, freight movements are concentrated along the main radial routes, as well as along the lines linking the industrial centres of northern and northeastern France.

Decline of rail freight traffic

Within an increasing number of urban areas, investment has been made in new underground rail and tram systems in an effort to reduce congestion on the roads and related problems of pollution. Provincial cities such as Lyon, Marseille, Lille, and Toulouse now boast metro networks, while a growing number of other cities (such as Lille, Nantes, Strasbourg, and Grenoble) are served by tramways, a solution increasingly favoured because of its comparatively lower cost. However, this has not stopped further substantial investment in the Paris Métro or the high-speed regional system (Réseau Express Régional; RER). Lines have been extended farther into the suburbs, and major new capacity has been added in central Paris.

Waterways. Despite the presence of major rivers such as the Seine, Rhine, and Rhône, inland waterways carry little freight. Although they are still used to transport goods

A high-speed TGV (*train à grande vitesse*) traversing the Burgundy region between Tournous and Mâcon.
© Riviere—Rapho/Photo Researchers

such as construction materials and agricultural and oil products, their role has progressively declined in the face of cheaper and faster alternatives. Traffic has also been lost because of the reduced inland movement of heavy raw materials and fuel products and an inefficiently organized industry with too many small-barge operators. The uneven and disjointed pattern of the waterways further restricts use. Less than one-third of the commercial waterway system is of European standard gauge; moreover, the principal river and canal systems remain unconnected for the passage of large barges, so that no truly national or international network exists.

France is served by a large number of maritime ports, which reflects not only its extensive coastline but also its importance as a trading nation. As in other Western countries, however, France's merchant fleet has steadily shrunk, largely because of the difficulty of competing with lower-cost carriers. Freight traffic, consisting mostly of imports, is concentrated in a limited number of ports, principally Marseille and Le Havre, followed by Dunkerque, Calais, Nantes-Saint-Nazaire, and Rouen. This imbalance is partly explained by the still-sizable quantities of crude oil that are unloaded. Passenger traffic is less important but is dominated by cross-channel movements from the port of Calais and the nearby Channel Tunnel.

Air transport. Air freight and passenger traffic have expanded rapidly and, like other forms of transport, are centred on Paris. The capital's two major airports (Roissy [Charles de Gaulle] and Orly) represent the second largest airport complex in western Europe (after London), handling roughly two-thirds of all French passenger traffic. Other French airports are far less important, though the country has a comprehensive network of local and regional airports. The majority of routes, however, are between provincial towns and cities and the capital rather than between regional centres, which reemphasizes the persistent centralization of economic activity and decision making in France. Nice and Marseille are the busiest regional air centres and, along with Lyon, Bordeaux, Toulouse, and Strasbourg, are the only provincial airports to have significant international traffic.

Telecommunications. At the beginning of the 21st century, France had some 34 million main telephone lines, almost all with digital capacity. Nevertheless, the nation lagged behind many EU countries in telecommunications in that just over half of its citizens used cellular telephones, only about one-third owned personal computers, and roughly one-sixth were Internet users. These comparatively low statistics were due in part to restrictive government controls on electronic commerce (e-commerce) and the presence of an existing network called Minitel (founded 1983 and owned by France Telecom)—obstacles that began to fall away in the first years of the 21st century. (J.N.T./Ed.)

Government and society

CONSTITUTIONAL FRAMEWORK

The genesis of the 1958 constitution. When France fell into political turmoil after the May 1958 insurrection in Algeria (then still a French colony), General Charles de Gaulle, an outspoken critic of the postwar constitution who had served as the provisional head of government in the mid-1940s, returned to political life as prime minister. He formed a government and, through the constitutional law of June 1958, was granted responsibility for drafting a new constitution. The drafting of the constitution of the Fifth Republic and its promulgation on Oct. 4, 1958, differed in three main ways from the former constitutions of 1875 (Third Republic) and 1946 (Fourth Republic): first, the parliament did not participate in its drafting, which was done by a government working party aided by a constitutional advisory committee and the Council of State; second, French overseas territories participated in the referendum that ratified it on Sept. 28, 1958; and, third, initial acceptance was widespread, unlike the 1946 constitution, which on first draft was rejected by popular referendum and then in a revised form was only narrowly approved. In contrast, the 1958 constitution was contested by 85 percent of the electorate, of which 79 percent were in favour; among the overseas territories only Guinea rejected the new constitution and consequently withdrew from the French Community.

Earlier constitutions

The dual executive system. In order to achieve the political stability that was lacking in the Third and the Fourth Republic, the constitution of 1958 adopted a mixed (semi-presidential) form of government, combining elements of both parliamentary and presidential systems. As a result, the parliament is a bicameral legislature composed of elected members of the National Assembly (lower house) and the Senate (upper house). The president is elected separately by direct universal suffrage and operates as head of state. The constitution gives the president the power to appoint the prime minister (often known as the premier), who oversees the execution of legislation. The president also appoints the Council of Ministers, or cabinet, which together with the prime minister is referred to as the government.

The role of the president. The French system is characterized by the strong role of the president of the republic. The office of the president is unique in that it has the authority to bypass the parliament by submitting referenda directly to the people and even to dissolve the parliament altogether. The president presides over the Council of Ministers and other high councils, signs the more important decrees, appoints high civil servants and judges, negotiates and ratifies treaties, and is commander in chief of the armed forces. Under exceptional circumstances, Article 16 allows for the concentration of all the powers of the state in the presidency. This article, enforced from April to September 1961 during the Algerian crisis, has received sharp criticism, having proved to be of limited practical value because of the stringent conditions attached to its operation.

De Gaulle's great influence and the pressures of unstable political conditions tended to reinforce the authority of the presidency at the expense of the rest of the government. Whereas the constitution (Article 20) charges the government to "determine and direct" the policy of the nation, de Gaulle arrogated to himself the right to take the more important decisions, particularly concerning foreign, military, and institutional policies, and his successors adopted a similar pattern of behaviour. The constitution of 1958 called for a presidential term of seven years, but, in a referendum in 2000, the term was shortened to five years, beginning with the 2002 elections.

The role of the prime minister, however, has gradually gained in stature. Constitutionally, the office is responsible for the determination of governmental policy and exercises control over the civil service and the armed forces. Moreover, while all major decisions tended to be taken at the Élysée Palace (the residence of the president) under de Gaulle, responsibility for policy, at least in internal matters, has slowly passed to the head of the government. Especially since the mid-1970s, a working partnership between the president and the prime minister has tended to be established. Finally, the power of the president is tied to the parliamentary strength of the parties that support him and that form a majority in the National Assembly. It is possible, however, for the president's parties to become a minority in the assembly, in which case the president must appoint a prime minister from the majority faction. From 1986 through the beginning of the 21st century, except for two years (1993–95), France experienced a form of divided government known as "cohabitation," in which the president and the prime minister belonged to different parties.

The National Assembly and the Senate

Parliamentary composition and functions. The National Assembly is composed of 577 deputies who are directly elected for a term of five years in single-member constituencies on the basis of a majority two-ballot system, which requires that a runoff take place if no candidate has obtained the absolute majority on the first ballot. The system was abandoned for proportional representation for the 1986 general election, but it was reintroduced for the 1988 election and has remained in place ever since. The Senate is composed of 321 senators indirectly elected for nine years by a *collège électoral* consisting mainly of municipal councillors in each *département*, one of the administrative units into which France is divided. The parliament retains its dual function of legislation and control over the executive but to a lesser extent than in the past. The domain of law (Article 34) is limited to determining the basic rules and fundamental principles concerning such matters as civil law, fiscal law, penal law, electoral law, civil liberties, labour laws, amnesty, and the budget. In these matters the parliament is sovereign, but the government can draw up the details for the application of laws.

The government is responsible for all other matters, according to Article 37 of the constitution, and the assemblies can in no way interfere; the Constitutional Council is responsible for ensuring that these provisions are respected. The parliament can temporarily delegate part of its legislative power to the government, which then legislates by ordinances. This procedure has been used on matters concerning Algeria, social security, natural disasters, European integration, and unemployment. Finally, government and the parliament are advised by an Economic and Social Council, composed of 230 representatives of various groups (*e.g.*, trade unions and employers' and farmers' organizations) that must be consulted on long-term programs and on developments and that may be consulted on any bill concerning economic and social matters.

The right to initiate legislation is shared by the government and the parliament. Bills are studied by parliamentary committees, although the government does control the agenda. The government can also, at any point during the debate over a bill, call for a single vote on the whole of the bill's text. Parliamentary control over the government can be exercised, but it is less intense than in the British system. There are questions to ministers challenging various aspects of performance, but these take place infrequently and are primarily occasions for lesser debates and do not lead to effective scrutiny of the government's practices. Committee inquiries are also relatively rare. The National Assembly, however, has the right to censure the government, but, in order to avoid the excesses that occurred before 1958 (as a result of which governments often fell once or twice a year), the motion of censure is subject to considerable restrictions. Only once in the first 50 years of the Fifth Republic, in 1962, did the National Assembly pass a motion of censure, when it stalled de Gaulle's referendum for direct election of the president by universal suffrage, which ultimately met with approval. The government is also strengthened by its constitutional power to ask for a vote of confidence on its general policy or on a bill. In the latter case a bill is considered adopted unless a motion of censure has obtained an absolute majority.

The motion of censure

The role of referenda. The people may be asked to ratify, by a constituent referendum (Article 89), an amendment already passed by the two houses of the parliament. The constitution made provision for legislative referenda, by which the president of the republic has the authority to submit to the people a proposed bill relating to the general organization of the state (Article 11).

This procedure was used twice in settling the Algerian question of independence, first in January 1961, to approve self-determination in Algeria (when 75 percent voted in favour), and again in April 1962, approving the Évian Agreement, which gave Algeria its independence from France (when 91 percent voted in favour). The use of this latter procedure to amend the constitution without going through the preliminary phase of obtaining parliamentary approval is constitutionally questionable, but it led to a significant result when, in October 1962, the election of the president by universal suffrage was approved by 62 percent of those voting. In April 1969, however, in a referendum concerning the transformation of the Senate into an economic and social council and the reform of the regional structure of France, fewer than half voted in favour, and this brought about President de Gaulle's resignation.

Through the end of the 20th century, national referenda were met with low voter turnout. The procedure was used in 1972 for the enlargement of the European Economic Community (EEC; later the European Community, which became part of the EU) by the proposed addition of Denmark, Ireland, Norway, and the United Kingdom; in 1988 for the proposed future status of the overseas territory of New Caledonia; and in 1992 for approval of the Maastricht Treaty, which established the EU. In 1995, when minor modifications were made to the constitution, the use of the referendum was enlarged to include proposed legislation relating to the country's economic and social life. In 2000 a referendum shortened the presidential term from seven to five years.

The role of the Constitutional Council. The Constitutional Council is appointed for nine years and is composed of nine members, three each appointed by the president, the National Assembly, and the Senate. It supervises the conduct of parliamentary and presidential elections, and it examines the constitutionality of organic laws (those fundamentally affecting the government) and rules of parliamentary procedure. The council is also consulted on international agreements, on disputes between the government and the parliament, and, above all, on the constitutionality of legislation. This power has increased over the years, and the council has been given a position comparable to that of the U.S. Supreme Court.

REGIONAL AND LOCAL GOVERNMENT

The units of local government are the *régions*, the *départements*, the *communes*, and the overseas territories.

Establishment of the *régions*

The *régions*. One of the main features of decentralization in French government has evolved through the creation of the 22 *régions*. After a number of limited changes lasting two decades, a 1982 law set up directly elected regional councils with the power to elect their executive. The law also devolved to the regional authorities many functions hitherto belonging to the central government, in particular economic and social development, regional planning, education, and cultural matters. The *régions* have gradually come to play a larger part in the administrative and political life of the country.

The *départements*. The *région* to an extent competes with the *département*, which was set up in 1790 and is still regarded by many as the main intermediate level of government. By the late 20th century, the initial number of 83 *départements* had reached 100: 96 in metropolitan France and 4 overseas (Guadeloupe, Martinique, French Guiana, and Réunion). Each *département* is run by the General Council, which is elected for six years with one councillor per *canton*. There are between 13 and 70 *cantons* per *département*. The General Council is responsible for all the main departmental services: welfare, health, administration, and departmental employment. It also has responsibility for local regulations, manages public and private property, and votes on the local budget. A law passed in 1982 enhanced decentralization by increasing the powers and authority of the *départements*. Formerly, the chief executive of the *département* was the government-appointed prefect (*préfet*), who also had strong powers over other local authorities. Since the law went into effect, however, the president of the General Council is the chief executive and the prefect is responsible only for preventing the actions of local authorities from going against national legislation.

The *communes*. The *commune*, the smallest unit of democracy in France, dates to the parishes of the ancien régime in the years before the Revolution. Its modern structure dates from a law of 1884, which stipulates that *communes* have municipal councils that are to be elected for six years, include at least nine members, and be responsible for "the affairs of the *commune*." The council administers public land, sets up public undertakings, votes on its own budget, and over recent years has played an increasing role in promoting local economic development. It elects a mayor and the mayor's assistants. Supervision by the central government, once very tight, has been markedly reduced, especially since 1982.

The importance of the mayor

The mayor is both the chief executive of the municipal council and the representative of the central government in the *commune*. The mayor is in charge of the municipal police and through them ensures public order, security, and health and guarantees the supervision of public places to prevent such things as fires, floods, and epidemics. The mayor also directs municipal employees, implements the budget, and is responsible for the registry office. French mayors are usually strong and often dominate the life of the *commune*. They are indeed important figures in the political life of the country.

French *communes* are typically quite small; there are more than 36,500 of them. Efforts have been made to group *communes* or to bring them closer to one another, but these have been only partly successful. In certain cities, such as Lyon and Lille, cooperative urban communities have been created to enable the joint management and planning of a range of municipal services, among them waste disposal, street cleaning, road building, and fire fighting. A similar approach has been adopted elsewhere, including rural areas, with the establishment of *syndicats intercommunaux* that allows services to be administered jointly by several *communes*. Moreover, since the 1999 law on Regional Planning and Sustainable Development, the *communes* within urban areas of more than 50,000 inhabitants have been encouraged to pool resources and responsibilities to promote joint development projects by means of a new form of administrative unit known as the *communauté d'agglomération*.

• **The overseas territories.** The status of many of France's overseas territories—vestiges of the French Empire—changed in the 1970s. Independence was proclaimed in 1975 by the Indian Ocean archipelago of the Comoros, with the exception of Mayotte (Mahoré) island, which chose to remain within French rule; in 1977 by Djibouti, on the Horn of Africa; and in 1980 by the Anglo-French Pacific Ocean condominium of the New Hebrides, under the name of Vanuatu. Mayotte was elevated to the status of territorial collectivity in 1976, and in North America the island territory of Saint-Pierre and Miquelon was elevated to the same status in 1985.

The only places retaining overseas territory status are French Polynesia (with its capital at Papeete on the island of Tahiti), New Caledonia, the Wallis and Futuna islands in the Pacific, and the Adélie Land claim in Antarctica. These territories have substantial autonomy except in matters reserved for metropolitan France, such as diplomacy and defense. They are governed through various but similar administrative structures, usually involving an elected council and a chief executive, but they are subject to the tutelage of a representative of the French Republic. A 1998 decision regarding New Caledonia envisaged the progressive transfer of political responsibilities to the island over a period of 15 to 20 years. Following decades of separatist violence, the overseas territory of Corsica was granted greater autonomy in the first years of the 21st century.

JUSTICE

In France there are two types of jurisdictions: the judiciary that judges trials between private persons and punishes infringements of the penal law and an administrative judicial system that is responsible for settling lawsuits between public bodies, such as the state, local bodies, and public establishments, as well as private individuals.

Civil courts

The judiciary. For civil cases the judiciary consists of higher courts (*grande instance*) and lower courts (*tribunaux d'instance*), which replaced justices of the peace in 1958. For criminal cases there are *tribunaux correctionnels* ("courts of correction") and *tribunaux de police*, or "police courts," which try minor offenses. The decisions of these courts can be referred to one of the 35 courts of appeal. Felonies are brought before the assize courts established in each *département*, consisting of three judges and nine jurors.

All these courts are subject to the control of the Court of Cassation, as are the specialized professional courts, such as courts for industrial conciliation, courts-martial, and, from 1963 to 1981, the Court of State Security, which tried felonies and misdemeanours against national security. Very exceptionally, in cases of high treason, a High Court of Justice (Cour de Justice de la République), composed of members of the National Assembly and of senators, is empowered to try the president of the republic and the ministers. They can also be tried by this court if they have committed felonies or misdemeanours during their term of office. These are the only situations in which the Court of Cassation is not competent to review the case. Otherwise, the court examines judgments in order to assess whether the law has been correctly interpreted; if it finds that this is not the case, it refers the case back to a lower court.

The more than 5,000 judges arc recruitcd by means of competitive examinations hcld by the National School of the Magistracy, which was founded in 1958 and in 1970 replaced the National Centre for Judicial Studies. A traditional distinction is made between thc *magistrats du siège*, who try cases, and the *magistrats de parquet* (public prosecutors), who prosecute. Only the former enjoy the constitutional guarantee of irremovability. The High Council of thc Judiciary is made up of 20 members originally appointed by the head of state from among the judiciary. Since 1993, however, its members have been elected, following reforms designed to free the judiciary from political control. The Council makes proposals and gives its opinion on the nomination of the *magistrats du siège*. It also acts as a disciplinary council. Public prosecutors act on behalf of thc statc. Thcy are hierarchically subject to the authority of the ministcr of justice. Judges can serve successively as members of the bench (*siège*) and the public prosecutor's department. They act in collaboration with, but are hierarchically independent of, the police.

Administrative courts. One of the special characteristics of the French judicial system is the existcnce of a hierarchy of administrative courts whose origins date to Napoleon. The duality of the judicial system has been sometimes regarded unfavourably, but the system has come to be gradually admired and indeed widely adopted in continental European countries and in the former French colonies. The administrative courts are undcr the control of the Council of State, which examines cases on appeal. The Council of State thus plays a crucial part in exercising control over the government and the administration from a jurisdictional point of view and ensures that they conform with the law. It is, moreover, empowered by the constitution to give its opinion on proposed bills and on certain decrees.

POLITICAL PROCESS

Universal suffrage at age 21 has existed in France since 1848 for men and since 1944 for women; the age of eligibility was lowered to 18 in 1974. Legislation enacted in the late 1990s penalizes political parties for failing to maintain sufficient parity between male and female candidates. Candidates for the National Assembly must receive a majority, not a plurality, of votes, and, if no candidate receives an absolute majority, then a second ballot is held the following week and the post is awarded to the plurality winner. Elections follow the model of single-member districts rather than proportional representation within a district. Two-phase voting is also used for the presidency, with the exception that, if an absolute majority is not reached after the first ballot, then only the two highest vote getters are considered for the second ballot, which is contested two weeks later.

Historically, French political parties have been both numerous and weak, which is generally accepted as the reason governments fell frequently before the advent of the Fifth Republic in 1958. Since then there has been a degree of streamlining, although, especially among centrist groups, parties are still poorly organized and highly personalized. Indeed, there have been many vicissitudes in the fortunes of the main parties since the late 1950s. In the 1960s and early '70s, Charles de Gaulle's centre-right party—first named Union for the New Republic (UNR) and later Rally for the Republic (RPR)—dominated the elections. After the election of the centrist Valéry Giscard d'Estaing to the presidency in 1974, the Gaullist party declined, while the centrists (from 1978 as the Union for French Democracy; UDF) and Socialists gained in strength. From 1981, and with the election of the Socialist president François Mitterrand, the Socialist Party became dominant, with its gains made primarily at the expense of the Communists. It was the first time since 1958 that the left had taken the leadership in French politics. While the Gaullists achieved a comeback with the election of Édouard Balladur as prime minister in 1993 and of Jacques Chirac as president in 1995, the Socialists regained control of the government in 1997 when Lionel Jospin became prime minister. In 2002 Chirac was reelected to the presidency under the coalition banner of the Union for Presidential Majority, decisively putting down Jean-Marie Le Pen of the far-right National Front, who surprised many with his strong showing in the first round of balloting. Chirac named Jean-Pierre Raffarin as his prime minister.

The French party system has continued to display volatility, though less so than in the past. Because the dominance of the Gaullist party was relatively short-lived, with other groups from the centre eroding its strength, the parliamentary base of the governments of the centre-right shrank; this was especially so since the centrists remained a loose confederation of several groupings, each of which tended to adopt different tactics. The precarious nature of political balance was underscored by recent periods of cohabitation between presidents and prime ministers of opposing parties.

SECURITY

Armed forces. The overall responsibility for national defense rests with the president, who is the constitutional chief of the armed services and presides over the higher councils and committees on national defense. Since a decree in 1964, the president can give the order to bring the air and strategic forces into action. The prime minister, assisted by the secretary-general for national defense, oversees the armed forces according to the terms of the constitution, but it is the minister of defense who actually directs the land, air, and naval forces and who, moreover, has authority over the armament policy and the arsenal.

Since 1958 the military administration has been divided by various functions; it includes strategic nuclear forces, territorial-defense forces, mobile forces, and task forces. France has had the atomic bomb since 1960 and the hydrogen bomb since 1968. The nation withdrew from the integrated command of the North Atlantic Treaty Organization (NATO) in 1966 but rejoined in 1995. An all-volunteer army was in place by 2002, though previously every French male 18 years of age had been subject to one year of compulsory military service.

Nuclear weapons

Police services. The police are responsible primarily for maintaining public law and order. Under the authority of the minister of interior, they are responsible to the prefects in the *départements* and to the prefect of police in Paris and adjacent suburban *communes*. The police force is divided into public security forces and specialized police forces, such as the vice squad. The security police include the State Security Police (Compagnies Républicaines de Sécurité; CRS), responsible for public order; the judicial police, who carry out criminal investigations and hunt down suspects; and the complex internal intelligence and antiespionage units. The municipal forces are responsible to the mayor. There is also the national gendarmerie, a kind of state police, which is responsible to the minister of defense, combats terrorism, and is of particular importance in the rural areas.

The gendarmerie

HEALTH AND WELFARE

Social security and health. Almost everyone is covered by the social security system, notably after the reform of 1998 that extended coverage to those previously excluded owing to lack of income. Social insurance was introduced in 1930 and family allowances in 1932, but the comprehensive rules for social security were established in 1946. A network of elected social security and family allowance *caisses primaires* ("primary boards"), headed by national *caisses*, manages a considerable budget. This budget relies on employers' and employees' social security contributions, as well as the proceeds from a special tax, introduced in 1991 (*contribution sociale généralisée*), on all forms of income. Deficits are made up by the state. The majority of expenditure is devoted to retirement benefits (pensions) and the partial reimbursement of most medical expenses. Other payments include family benefits for dependent children, unemployment indemnities, and housing subsidies. Since 1988, in response to a long-term problem of unemployment in France, people with little or no income have been able to benefit from a special government subsidy known as the social minimum (*revenu minimum d'insertion*).

France complies with the principles of liberal medicine, with patients free to choose doctors and treatment. Since 1960, however, agreements have been signed at a regional level between the *caisses* and the professional medical associations that regulate fees. Although doctors need not necessarily adhere to them, reimbursements from social security are based on these scales.

Hospital reforms

The hospital reform of 1960 joined hospitals and medical schools through the creation of teaching hospitals. Private hospitals and clinics operate alongside public hospitals, and the cost of treatment in private facilities may also be partially reimbursed from social security funds. Since the enactment of legislation in 1991, the government has sought to rationalize the distribution of hospitals to take advantage of shifting population densities, changing health care needs, and new technology.

Housing. The government encourages construction through premiums, loans (particularly for low-rent housing), and tax incentives. Municipal and other public bodies also have engaged in a vast program of subsidized public housing (*habitation à loyer modéré*; HLM), which was especially prominent in the 1960s and '70s. In 1970 the procedure for receiving building permits for private construction was greatly simplified, and since 1982 mayors have been responsible for granting construction permits and devising local housing policies for both the public and private sectors. The government has also sought to encourage home ownership through low-interest loans. As a result of continuing suburbanization, far greater emphasis is now placed on building houses rather than apartments. From the late 1960s city planning in France became more organized through such programs as *zone d'aménagement concerté* (ZAC), which often link both private and public developers. Reforms in 2000 updated long-term development plans (*schéma de cohérence territoriale*; SCOT) and detailed land-use plans (*plan local d'urbanisme*; PLU). The current emphasis of urban policy is on rehabilitation, particularly of the many peripheral housing estates built in the 1960s and '70s but also of older central districts.

Wages and the cost of living. Despite a history of high inflation, over recent years levels have been similar to those of other industrialized countries. Indeed, since the mid-1980s inflation has been particularly low in France. A minimum wage law has been in effect since 1950, and since 1970 it has been supplemented by a provision known as the *salaire minimum interprofessionel de croissance* (SMIC; general and growth-indexed minimum wage), which has increased the lowest salaries faster than the inflation rate. Its level is set annually, and all employers must abide by it. Women are, in general, paid less well than men. A worker earns nearly twice as much in Paris as in the less-developed *départements* of central France. The differentials between the earnings of manual workers and those of managers, while still large, have diminished progressively. In general, the majority of French society has benefited from a very substantial increase in purchasing power over the last half century.

EDUCATION

The organization of national education is highly centralized. Since 1968, however, following rioting among university students seeking a greater voice in their administration, a movement toward decentralization has been in progress in higher education. Reforms have sought to modify the character and structure of education, not only at the university level but also in primary and secondary schools; in the latter case one of the principal government aims has been to enable 80 percent of secondary-school students to obtain their *baccalauréat*.

France has both public and private education. All public education is free and is administered by the Ministry of National Education, which draws up the curricula, employs the staff, and exercises its authority through rectors placed at the heads of academies. However, while the state retains control of the educational programs and faculty, responsibility for the provision and maintenance of schools has been decentralized since the early 1980s; the *communes* look after primary schools, while in secondary education the *départements* are responsible for the *collèges* and the *régions* maintain the *lycées*.

Primary and secondary education. Education is free and compulsory between ages 6 and 16. Children under age 6 can attend *écoles maternelles* (nursery schools). Primary schools provide elementary education for those between ages 6 and 11. Secondary education begins in the *collèges* from ages 11 to 15 with further secondary education offered in general or technical *lycées*, leading to the national *baccalauréat* examination. Courses of study lasting for two or three years can lead to professional certificates or diplomas. School councils allow teachers and representatives of parents (and pupils at the secondary level) to gather to discuss the operation of schools.

Decentralization

Higher education. Following the student riots of May 1968, higher education was profoundly changed with the enactment of reforms on Nov. 12, 1968, though many centralizing features from the past remain. Previously, universities had been divided into faculties or colleges according to the subjects taught. After 1968 the faculties were replaced by teaching and research units regrouped into autonomous multidisciplinary universities comanaged by representatives elected from among the teaching staff, students, and administration. These institutions substantially determine their own research programs, teaching methods, and means of assessment. Much of the curriculum is still validated at the national level, however.

The state grants funds to the universities, which they divide among their departments. The degrees awarded are the *licence* (roughly comparable to the British-American bachelor's degree), *maîtrise* (master's degree), and doctorate. There are also special teaching qualifications, one of which is the *agrégation*, a rigorous competitive examination. Traditional university courses were considerably diversified by the creation of specialized technological sections (Instituts Universitaires de Technologies; IUT) in 1966 and by the establishment in 1991 of vocational units (Instituts Universitaires Professionnalisés; IUP), which work closely with businesses. Students may also apply to a number of prestigious *grandes écoles*, which are even more highly regarded than the universities, especially in the engineering and technical fields. The best-known among these is the École Polytechnique ("Polytechnic School"); founded in 1794 to recruit and train technicians for the army, it has become the most important technical school in both the public and private sectors.

Other features. Private education is mostly Roman Catholic. Although the French constitution proclaims that the state is secular, a 1959 law allows private establishments to sign government contracts that procure financial support in exchange for some control. Despite attempts made by the Socialist government of the early 1980s to bring private schools closer to the public sector, the system has remained basically unchanged.

Teachers are highly unionized and belong largely to the Federation for National Education and the National Syndicate of Instructors, as well as other left-wing unions. The main student unions are the National Union of French Students (Union National des Étudiants de France; UNEF), the quasi-communist UNEF Renouveau, the Union of Communist Students of France, and the National Confederation of French Students.

(F.B./J.F.P.B./J.N.T./Ed.)

Cultural life

For much of its history, France has played a central role in European culture. With the advent of colonialism and global trade, France reached a worldwide market, and French artistic, culinary, and sartorial styles influenced the high and popular cultures of countries around the globe. Today French customs, styles, and theories remain an influential export, as well as a point of great national pride, even as French intellectuals worry that the rise of globalism has prompted, in the words of the historian Pierre Nora, "the rapid disappearance of our national memory."

CULTURAL MILIEU

French culture is derived from an ancient civilization composed of a complex mix of Celtic, Greco-Roman, and Germanic elements. Monuments, especially from the period of Roman occupation, are numerous and include the amphitheatre at Arles, the *arènes* ("arenas") in Paris, and the aqueduct at Pont du Gard.

Cultural development in the Middle Ages

During the Middle Ages a rich culture developed, fostered in particular by monks and scholars in monasteries and universities and encouraged well into the 18th century by a system of royal and aristocratic patronage. Important trade fairs in growing cities such as Paris, Nancy, Strasbourg, and Lyon enabled the spread of artistic ideas and cultural trends to and from other regions, placing France at the centre of a nascent European high culture that would reach its greatest expression in the Renaissance. From the early 1700s and with the development of a middle class, the bourgeoisie, culture became more generally accessible. This was the age of the Enlightenment, of inquiry and question. Cultural activity remained largely centred in Paris, but smaller cities such as Aix-les-Bains, Grenoble, and Lyon were vital in their own right. The culture of the Enlightenment was built on reason and analytic argumentation, mirrored, as political scientist Alexis de Tocqueville remarked, in the French Revolution's

> attraction for general theories, for general systems of legislation, the exact symmetry of laws . . . the same desire to remake the entire constitution at once following the rules of logic and in accordance with a single plan, instead of seeking ways to amend its parts.

Among its tenets was the idea of meritocracy, or an aristocracy of ability and intelligence, which accorded a central place to intellectuals unknown in most other societies and opened France's schools to students from the provinces without regard for social class.

With free primary education compulsory by the late 19th century, basic literacy ensured that the general cultural level was raised. This was further aided by the increase in the number of newspapers and, later, by the development of radio, cinema, television, and the Internet. After World War II the intellectual and social development of lower-income groups benefited from the decision to make free secondary education compulsory up to age 16. Cultural literacy expanded as newspaper circulations rose, lending libraries proliferated, and in 1954 a revolution began in paperback books (*livre de poche*). This last development met with enormous success, providing people of all ages and classes with much greater access to literature and other forms of specialized knowledge.

The Ministry of Culture and Communications oversees the major cultural institutions of the nation. The department, first led by novelist André Malraux, seeks to redouble arts awareness among ordinary people, support the creation of new art, and protect existing French forms and properties as wide-ranging as monuments and language. The cultural map of France remains firmly centred on Paris, despite increased expenditure by local authorities on cultural activities following the decentralization legislation of the early 1980s. Yet, while serving, often self-consciously, the interests of the whole nation, the capital is aware of its own internal differences. Most of the city's *arrondissements* (municipal districts) have groups actively researching their history and traditions, and local art exhibitions and concerts are encouraged. In the rest of the country, provincial culture is strong and often fiercely defended—for example, in Brittany, parts of the south, and Alsace.

French culture has felt the impact made by immigrants, especially those from North Africa beginning in the 1960s. The Muslim communities that have formed, notably in Paris and Marseille, have not escaped discrimination, but there is a widespread acknowledgment of their contributions to cuisine, music, dance, painting, and literature. Verlan, a slang of standard French that reverses and reshuffles French syllables and spellings, traces its roots to the 19th century but was revived by postwar immigrant communities and in recent decades has made inroads into mainstream society. Beginning in the 1980s, second- and third-generation North Africans were often referred to as *les beurs*, and *beur* cinema, *beur* comics, and *beur* radio, among other forms of expression, have found a large audience. The label *beur* is itself a Verlan term for *arabe*, the French word for Arab. In addition, Asian and sub-Saharan African immigrants have attained prominence as artists, writers, and musicians in France's increasingly multicultural society.

DAILY LIFE AND SOCIAL CUSTOMS

In comparison with the immediate postwar era, the French now devote far more time to leisure and cultural pursuits, largely as a result of a shorter workweek, more years spent in education, and greater affluence. The increasing emphasis on home entertainment provided by television, stereo, and personal computers has not reduced cinema or theatre attendance; on the contrary, the number of moviegoers grew significantly in the 1990s.

The popularity of cultural activities is also evident, with increasing visits to historic monuments, art galleries, and museums. Especially attractive are interactive exhibitions at museums, such as the Cité de Sciences et de l'Industrie (City of Science and of Industry) at Le Parc de la Villette in Paris or the Futuroscope theme park near Poitiers. Interest also has been revived in local and regional cultures, often as part of new initiatives to develop tourism, and annual national festivals, such as the Fête de la Musique, are extremely successful.

French cuisine

Although French cuisine has a reputation as a grand national feature, regional differences are marked. Some local dishes have achieved international fame, even if they are often poorly imitated. Among these are the seafood soup, *bouillabaisse*, from Marseille; *andouillette*, a form of sausage from Lyon; *choucroute*, pickled cabbage from Alsace; and *magret de canard*, slices of breast of duck from Bordeaux. France is also renowned for the range and quality of its cheeses. More than 300 varieties are recognized. The majority are produced from cow's milk, including Camembert (Normandy), Brie (Île-de-France), Comté (Franche-Comté), Saint-Nectaire (Auvergne), and Reblochon (Savoy). Cheese is also made from ewe's milk, as in the case of Roquefort (Aveyron), as well as from goat's milk. Perhaps the best-known exports of France are the wines from some of the world's great vineyards in Burgundy, Bordeaux, and the Rhône valley. However, the reputation of French cuisine has not prevented the proliferation of fast-food outlets in France, especially over the past few decades. French consumption of wine and tobacco has dropped steadily since the mid-20th century, a mark of the nation's increased attention to health.

Paris is internationally known for its haute couture, exemplified by such houses of high fashion as Coco Chanel, Christian Dior, Hubert de Givenchy, Yves Saint Laurent, Jean-Paul Gaultier, and Christian Lacroix. Traditional dress is occasionally seen in many regions, although it is largely reserved for official ceremonies and festivals. Regional differences often reflect local customs of dressmaking and embroidery, the availability of fabrics, and adaptations to local climatic conditions. Headdresses vary greatly, ranging from elaborate lace wimples found in Normandy and Brittany to the more sober beret of southwestern France or the straw hat, worn typically in and around the area of Nice.

Mal Langsdon—© Reuters/Corbis

Jets trailing the French national colours over the Champs-Élysées in Paris, during the annual Bastille Day parade.

In addition to the Roman Catholic holy days, the French celebrate Bastille Day on June 14, commemorating the rise of the French Republic via the fall of the prison fortress of the Bastille in Paris in 1789 at the start of the French Revolution. "La Marseillaise," the French national anthem and one of the world's most recognizable national anthems, also memorializes the Revolution.

THE ARTS

Literature. French literature has a long and rich history. Traditionally it is held to have begun in 842 with the Oath of Strasbourg, a political pact between Louis the German and Charles the Bald, the text of which survives in Old French. The Middle Ages are noted in particular for epic poems such as *La Chanson de Roland* (*c.* 1100; *The Song of Roland*), the Arthurian romances of Chrétien de Troyes, and lyric poetry expressing romantic love. In the 16th century the Renaissance flourished, and figures such as the poet Pierre de Ronsard, the satirist and humorist François Rabelais, and the quintessential essayist Michel de Montaigne, were to become internationally acknowledged. French Neoclassical drama reached its apotheosis during the next hundred years in the tragedies of Pierre Corneille and Jean Racine. During the same period, Molière displayed his vast and varied talents in the theatre, particularly as a writer of comedies; Jean de La Fontaine produced moralistic verse in his *Fables*; and Madame de La Fayette created the classic *La Princesse de Clèves* (1678), generally considered the first French psychological novel.

Voltaire, Denis Diderot, and Jean-Jacques Rousseau dominated the 18th century, especially with their philosophical writings, though they made major contributions to all genres, and Voltaire's novel *Candide* (1759) is notable for its literary quality and distillation of Enlightenment ideals. Other authors of the period include playwright Pierre-Augustin Caron de Beaumarchais, best known for works such as *Le Mariage de Figaro* (1784; *The Marriage of Figaro*), and Pierre Choderlos de Laclos, remembered for his epistolary novel *Les Liaisons dangereuses* (1782; *Dangerous Acquaintances*). The 19th century witnessed the emergence of a series of writers who substantially influenced the development of literature worldwide, including the novelists Honoré de Balzac, Stendhal, Gustave Flaubert, and Émile Zola along with the poets Charles Baudelaire, Stéphane Mallarmé, and Arthur Rimbaud. Added to these was the Romantic writer Victor Hugo, whose creative energy expressed itself in all literary forms, as well as in painting.

The "new novel"

French literature in the 20th century both carried on the earlier traditions and transformed them. While the complexity of French poetry continued in the work of Paul Valéry, Guillaume Apollinaire, Paul Claudel (also a major dramatist), Saint-John Perse, Paul Éluard, Louis Aragon, René Char, and Yves Bonnefoy, the art of the novel was given new direction by Marcel Proust, in *À la recherche du temps perdu* (1913–27; *Remembrance of Things Past*). The first half of the century also produced such notable writers as André Gide, François Mauriac, André Malraux, Albert Camus, and Jean-Paul Sartre, the last arguably the chief exponent of existentialist philosophy. Their work was followed in the 1950s by the *nouveau roman* ("new novel") and by the emergence of writers such as Alain Robbe-Grillet, Nathalie Sarraute, Michel Butor, Claude Mauriac, Marguerite Duras, and Claude Simon, whose works have entered the canon of literature. Since the 1970s Michel Tournier, Patrick Modiano, Erik Orsenna, and Georges Perec have become leading novelists; feminist writers, including Hélène Cixous, Annie Leclerc, Jeanne Champion, and Marie Cardinal, have also made significant contributions. French authors have won a number of Nobel Prizes for Literature.

The literature of the 20th century was notable for its openness to nonnative writers: the Irish writer Samuel Beckett, for instance, the Czech expatriate Milan Kundera, the Russian emigrant Andreï Makine, and Chinese exile Gao Xingjian have all produced major works in French. Georges Simenon and Marguerite Yourcenar, both born in Belgium in 1903, were considered French writers, though they often lived outside France.

The works of French playwrights have enjoyed international acclaim for centuries, from the 17th-century comic theatre of Molière to the 19th-century cabaret productions known as Grand Guignol. In theatre in the 20th and 21st centuries three important currents can be discerned. Traditional playwriting was carried on largely by Jean Anouilh, Claudel, Jean Giraudoux, Henry de Montherlant, and Camus, but experimentation with both form and content also developed. Before World War II, Jean Cocteau in particular made his mark (as did to a lesser degree Claudel), but innovation came with Fernando Arrabal, Arthur Adamov, Beckett, Jean Genet, and the Romanian exile Eugène Ionesco. Since the 1950s producers have also made an important contribution to theatre; Roger Planchon, Jean-Louis Barrault, Peter Brook, Marcel Maréchal, and Ariane Mnouchkine in particular have shared in both creating new works and revitalizing traditional ones.

Philosophy and criticism have always played a central part in French intellectual and cultural life. The Surrealist movement, led by André Breton, among others, flourished in the 1920s and '30s. Existentialism in both Christian and atheist forms was a powerful force in the mid-20th century and was championed by Sartre, Étienne Gilson, Gabriel Marcel, and Camus (though he rejected the label). More broadly, Roman Catholicism and Marxism in orthodox or revised forms have influenced a large number of creative writers, including the Roman Catholic Georges Bernanos and Sartre, who was a Marxist of a sort. Since the 1950s new criticism, which began with structuralism—itself largely inspired by the anthropological work of Claude Lévi-Strauss in *Mythologiques*, 4 vol. (1964–71), and *Tristes Tropiques* (1955)—has challenged the monopoly of the historical approach to works of art and especially literature. Not limited to literary criticism, structuralism was an important component of philosophy among proponents such as Louis Althusser. The most popular expression of this approach was perhaps the work of Roland Barthes, including *Mythologies* (1957), but his work fragmented into various branches—linguistic, genetic, psychobiographical, sociocultural—each with its exponents and disciples increasingly embroiled in academic, and often abstruse, debate. Following on the heels of structuralism, poststructuralism was associated with such figures as Jacques Derrida, Michel Foucault, Jacques Lacan, Julia Kristeva, Giles Deleuze, Hélène Cixous, Luce Irigaray, and Jean-François Lyotard. Other philosophers of recent note include André Glucksmann, Bernard Henri-Lévy, and Michel Serres.

The fine arts. French traditions in the fine arts are deep and rich, and painting, sculpture, music, dance, architecture, photography, and film all flourish under state support.

Painting and sculpture. In painting there was a long tradition from the Middle Ages and Renaissance that, while perhaps not matching those of Italy or the Low Countries, produced a number of religious subjccts and court portraits. By the 17th century, paintings of peasants by Louis Le Nain, of allegories and Classical myths by Nicolas Poussin, and of formally pastoral scenes by Claude Lorrain bcgan to givc Frcnch art its own charactcristics.

Within the next hundred years, styles bccamc evcn more wide-ranging: mildly erotic works by François Boucher and Jcan-Honoré Fragonard; cnigmatic sccnes such as *Pierrot*, or *Gilles* (*c.* 1718–19), by Antoinc Wattcau; interiors by Jean-Siméon Chardin that were oftcn tingcd with violence, as in *La Raie* (*c.* 1725–26; "The Ray"); emotive portraits by Jean-Baptiste Greuze; and rigorous Ncoclassical works by Jacques-Louis David.

Much as the Académie Françaisc rcgulatcd literature, painting up to this time was subjcct to rulcs and convcntions established by the Academy of Fine Arts. In thc 19th century some artists, notably J.-A.-D. Ingrcs, followcd these rulcs. Others reacted against academic conventions, making Paris, as the century progrcsscd, a ccntre of the Western avant-garde. Beginning in the 1820s, the bold croticism and "Orientalism" of the works of Romantic painter Eugène Delacroix angcrcd the academy, while at midcentury the gritty Realism of thc art of Gustavc Courbet and Honoré Daumier was viewed as scandalous.

The Imprcssionists

Perhaps the greatest brcak from academic conventions came about through the Impressionists, who, inspircd in part by the daring work of Édouard Manet, brought on a revolution in painting beginning in the late 1860s. Some artists from this movement whose work became internationally celebrated include Claude Monet, Camille Pissarro, Alfred Sisley, and Edgar Degas. Important Post-Impressionists include Paul Cézanne, Henri de Toulousc-Lautrec, Paul Gauguin, Pierre-Auguste Renoir, and Georges Seurat.

French sculpture progressed from the straight-lined Romanesque style through various periods to reach its height in the work of Auguste Rodin, who was a contemporary of the Impressionists and whose sculpture reflected Impressionist principles. Another from this time, Aristide Maillol, produced figures in a more Classical style.

Pablo Picasso, one of the most influential forces in 20th-century art, was born in Spain but spent most of his artistic life in France. His oeuvre encompasses several genres, including sculpture, but he is best known for the Cubist paintings he created together with French artist Georges Braque at the beginning of the century. One of Picasso's greatest rivals was French painter Henri Matisse, whose lyrical work, like Picasso's, spanned the first half of the century. In the period between the World Wars, Paris remained a major centre of avant-garde activity, and branches of prominent international movements such as Dada and Surrealism were active there.

By mid-century, however, Paris's dominance waned, and the focus of contemporary art shifted to New York City. Prominent artists working in France have included Jean Dubuffet, Yves Klein, Swiss-born Jean Tinguely, Hungarian-born Victor Vasarely, Niki de Saint-Phalle, Bulgarian-born Christo, Daniel Buren, and César.

Major art exhibits are held regularly, mainly in Paris, and training for aspiring artists is provided not only at the prestigious École des Beaux-Arts ("School of Fine Arts") in Paris but also at a number of provincial colleges. Courses for art historians and restorators are available at the School of the Louvre. Building on their country's rich history as a leader in furniture design and cabinetry, French craftsmen of all sorts today study at the National Advanced School of Decorative Arts and other institutions.

Music. The growth of classical music parallels that of painting. Despite work from earlier periods by Louis Couperin, Jean-Philippe Rameau, and Jean-Baptiste Lully, for example, French music gained a broad international following only in the 19th and early 20th centuries. Such composers as Hector Berlioz, Camille Saint-Saëns, Maurice Ravel, Claude Debussy, and the Polish-born Frédéric Chopin created a distinctively French style, further developed in the 20th century by composers such as Pierre Boulez, Darius Milhaud, and Erik Satie. In the late 20th century much experimentation occurred with electronic music and acoustics. The Institute for Experimentation and Research in Music and Acoustics (Institut de Recherche et Coordination Acoustique/Musique), in Paris, remains devoted to musical innovation. A new generation of French musicians includes the pianists Hélène Mercier and Brigitte Engerer.

Training for the musical profession remains traditional. Local conservatoires throughout the country provide basic grounding; some provincial schools—at Lyon and Strasbourg, for example—offer more advanced work, but young people with talent aim for the National Conservatory of Music in Paris, where Nadia Boulanger taught. Since World War II, Paris has hosted internationally famous conductors, such as Herbert von Karajan and Daniel Barenboim, who have made contributions in revitalizing an interest in classical music. Major visiting orchestras perform at the Châtelet Theatre or the Pleyel Concert Hall, and concerts are given by smaller groups in many of the churches. There is a network of provincial orchestras.

Although interest in classical music has grown at the amateur level, it is practiced by a relatively small number. The young tend to be preoccupied with popular music, especially that imported from the United States and the United Kingdom. The tradition of the French chanson, the romantic French ballad, has continued, however, following such legendary stylists as Juliette Gréco, Edith Piaf, Belgian-born Jacques Brel, Charles Aznavour, and Georges Brassens. Moreover, France has produced rock performers such as Johnny Hallyday and the group Téléphone, as well as chanteuses of the 1960s such as Françoise Hardy, known for pop music called *yé-yé* ("yeah-yeah"). Other well-known artists of recent years include Julien Clerc, Jean-Jacques Goldman, and Renaud. However, all have been considerably more popular nationally than internationally.

The French chanson

The Paris Opéra, established in 1669, prospered under the efforts of Lully, Rameau, Christoph Gluck, Berlioz, Georges Bizet, and Francis Poulenc. France was known for the traditions of opéra comique and grand opera, among others.

Dance. France is famous for developing ballet. In 1581 the *Ballet comique de la reine* was performed at the French court of Catherine de Médicis. Because it fused the elements of music, dance, plot, and design into a dramatic whole, it is considered the first ballet. The *Ballet comique* influenced the development of the 17th-century *ballet de cour* (court ballet), an extravagant form of court entertainment.

In 1661 Louis XIV established the Académie Royale de Danse (now the Paris Opéra Ballet); the company dominated European theatrical dance of the 18th and early 19th centuries. Pierre Beauchamp, the company's first director, codified the five basic ballet positions. Extending the range of dance steps were virtuosos such as Gaëtan Vestris and his son Auguste Vestris and also Marie Camargo, whose rival Marie Sallé was known for her expressive style.

In his revolutionary treatise, *Lettres sur la danse et sur les ballets* (1760), Jean-Georges Noverre brought about major reforms in ballet production, stressing the importance of dramatic motivation, which he called *ballet d'action*, and decrying overemphasis on technical virtuosity. In 1832 the Paris Opéra Ballet initiated the era of Romantic ballet by presenting Italian Filippo Taglioni's *La Sylphide*. Jean Coralli was the Opéra's ballet master at the time, and the company's dancers of this period included Jules Perrot and Arthur Saint-Léon.

In the 20th century ballet was rejuvenated under the leadership of Russian impresario Sergey Diaghilev, who founded the avant-garde Ballets Russes in Paris in 1909. For the next two decades it was the leading ballet company in the West. The original company was choreographed by Michel Fokine. Elsewhere in Paris, Serge Lifar, the Russian-born ballet master of the Paris Opéra, reestablished its reputation as a premier ballet troupe.

Dance entertainments of a lighter kind also were developed in France. In 19th-century Paris the all-female cancan became the rage. After 1844 it became a feature of music halls, revues, and operetta.

Architecture. With a rich and varied architectural heritage (which helped to spawn, among other styles, Gothic, Beaux Arts, and Art Deco) and an organized and competitive program of study, France has shown itself to be open to a variety of styles and innovations. For example, Le Corbusier, much of whose work can be found in France, was Swiss. The development of architecture has also been sustained by the central government's penchant for *grands projets*, or great projects. The country, however, has not produced as many designers of international repute in recent years as have other Western nations. Major achievements such as the Pompidou Centre, the pyramid entrance to the Louvre, and the Grand Arch have resulted from plans submitted in open competition by foreign architects. Recent architects of acclaim, of French origin or working in France, have included Jean Nouvel, Dominique Perrault, Adrien Fainsilber, Paul Andreu, Swiss-born Bernard Tschumi, and Catalonian Ricardo Boffil of Spain. Among important contemporary designers are Andrée Putman and Philippe Starck.

Photography. Jacques Daguerre, one of the recognized founders of modern photography in the early 19th century, began the evolution of an art form that has flourished in France. In the 20th century the work of such photographers as Eugène Atget, Henri Cartier-Bresson, and Robert Doisneau ensured that the art had a dimension beyond journalistic and commercial purposes, which was apparent in the installation art of later figures such as Christian Boltanski. In 1969 an annual festival was established at Arles, and in 1976 a national museum was created. The French popularization of photography through posters and postcards was one of the most remarkable cultural events of the late 20th century.

The cinema. French cinema has occupied an important place in national culture for more than 100 years. August and Louis Lumière invented a motion-picture technology in the late 19th century, and Alice Guy-Blaché and others were industry pioneers. In the 1920s French film became famous for its poetic realist mode, exemplified by the grand historical epics of Abel Gance and the work in the 1930s and '40s of Marcel Pagnol and others. A generation later the *nouvelle vague*, or New Wave, produced directors such as Jean-Luc Godard and François Truffaut, who "wrote" with the camera as if, in critic André Bazin's words, it were a *caméra-stylo* ("camera-pen"). This shift was accompanied by an "intellectualization" of the cinema reflected in the influential review *Cahiers du cinéma*, in the establishment of several schools in Paris and the provinces where film could be studied, and in the founding of film museums such as the Cinémathèque ("Film Library") in Paris.

Other directors of international stature include Jean Renoir, Jacques Tati, Jean-Pierre Melville, Alain Resnais, Eric Rohmer, Robert Bresson, and Louis Malle. They exemplified the auteur theory that a director could so control a film that his or her direction approximated authorship. Filmmakers such as Agnès Varda, Claude Chabrol, Jacques Demy, Bertrand Tavernier, and Claude Bérri, as well as Polish-born Krzystof Kieslowski, extended these traditions to the end of the century.

The leading film stars of the 20th century ranged from Fernandel, Maurice Chevalier, and Arletty to Brigitte Bardot, Gérard Depardieu, and Catherine Deneuve. One of the world's premier film festivals is held annually at Cannes, where the Palme d'Or is awarded to the best motion picture—most, in recent years, have come from outside France, a source of consternation to French film devotees. As in television, the French film industry faces competition from the United States and the United Kingdom. This led the government in the early 1990s to elicit the support of the European Commission to protect its native film industry.

CULTURAL INSTITUTIONS

Administrative bodies. Despite increasing support from the private sector, various ministries, such as those of National Education and of Culture and Communications, are ultimately responsible for the promotion of cultural activities. Local authorities, particularly those representing the major towns and cities, as well as a variety of associations also fund cultural activities. The importance attached to culture is reflected in the substantial increase in expenditure and personnel working in this field and the growth of related industries (music, publishing, broadcasting technologies). About one-third of the populace belong to some form of cultural association. Abroad, French culture is promoted through the work of counselors and attachés at embassies, visiting speakers, the Alliance Française, and the French lycées in major cities. French institutes provide lectures, language courses, and access to books and newspapers. There are also associations ensuring international links, such as the United Nations Educational, Scientific and Cultural Organization and the International Association of French Teachers, both headquartered in Paris.

French culture abroad

Museums and monuments. Support and encouragement for cultural activities of all kinds are provided by a large number of museums, centres, and galleries, many of which are ultimately the responsibility of government ministries. In the provinces many museums traditionally reflecting their region's activities have been expanded and renovated and, like those at Saint-Étienne and Strasbourg, have achieved national importance. It is in Paris, however, that the nation's principal museums are to be found. The Louvre Museum, containing one of the world's great art collections, was extensively remodeled at the end of the 20th century, with a notable addition of a dramatic steel-and-glass pyramid entrance. The Musée d'Orsay, created out of a former railway station, houses a fine, large collection of 19th- and early 20th-century art and artifacts, while the Georges Pompidou National Centre of Art and Culture, with its industrially inspired architecture, concentrates on the 20th century. The centre has an important library and media collection, and the square in front of it provides an open-air stage for jugglers, musicians, fire-eaters, and other street performers. Smaller museums, often containing substantial private collections, are numerous; three of particular interest are the Marmottan, Cognacq-Jay, and Orangerie. In addition to the larger museums, the Grand Palais and the Petit Palais regularly provide the setting for important exhibitions, and many of the national institutes offer French people the opportunity to appreciate works from different cultures. Particularly important in this respect are the Museum of the Arab World and the Museum of African and Oceanic Arts.

Since the 1950s there has been a national program for the conservation and renovation of important historic areas. The medieval *vieux quartiers* of Lyon have been tastefully restored, as has the 18th-century Place du Parlement in Bordeaux, for example. Many significant buildings have been saved by private funding, and government financial assistance is also available, usually on the condition that the property is opened to the public. In Paris the houses in the Marais district and on the Île Saint-Louis have had their original splendour restored, while around Montparnasse, for example, poor areas of 19th-century building have been bulldozed to make room for fashionable modern apartment blocks. Four structures in particular mark the later years of the 20th century: the entrance to the Louvre; the Bastille Opera; the Grand Arch in La Défense, a futuristic business district west of Paris; and the national library, Bibliothèque François Mitterrand, all of which received the strong support of Mitterrand as monuments to his presidency.

Significant structures of the late 20th century

SPORTS AND RECREATION

Although the French have recently developed a taste for a new range of sporting activities, such as mountain biking, cross-country skiing, and rock climbing, the most common forms of recreation in France seem to be nonphysical or relatively sedentary—talking, reading, eating, going to the cinema, and so on. This no doubt has something to do with the relative absence of programmed physical education at school. Certainly organized sport has a place in French society, however, with cycling, swimming, football (soccer), skiing, tennis, *boules* (*pétanque*), and, increasingly, golf, basketball, and martial arts being the most popular activities. Walking and jogging, too, are important, and a national network of paths (*grandes randonnées*) is well maintained. Popular seaside vacation resorts include Saint-Tropez, Cannes, and Cap d'Agde on the Mediterranean, the Île de Ré and La Baule-Escoublac on the Atlantic coast, and Le Touquet on the English Channel. Inland the French Alps, the Massif Central, and the national and regional parks, such as the Morvan regional nature park in Burgundy, attract campers and hikers. Newer, artificially created attractions include a growing number of theme parks, ranging from Disneyland at Paris to more specialized sites such as the Nausicaä sea-world museum at Boulogne-sur-Mer.

The nation's showcase sporting event is the Tour de France, an international cycling road race that attracts hundreds of thousands of spectators each year. Established more than a century ago, the annual summer race covers some 3,600 kilometres (2,235 miles) over the course of three weeks, finishing in Paris. Football, especially in the larger towns, is extremely popular. The 1998 World Cup was both hosted and won by France. More than five million French people ski, and many children have the opportunity to go on school skiing trips in February; the principal resorts are in the northern Alps, notably in Savoy (Savoie). French bowls, or *boules*, is played by thousands and is highly organized at both national and local levels. Handball has an avid following, and rugby is mostly played in the southwest. Educator Pierre, baron de Coubertin, revived the Olympic Games in modern form in 1896 and founded the International Olympic Committee. Games in Paris soon followed in 1900 and 1924. Chamonix was the site of the inaugural Winter Olympic Games in 1924, followed by Grenoble in 1960 and Albertville in 1992. Olympic highlights include the successes of skier Jean-Claude Killy in 1968, the national football team in 1984, and runner Marie-José Pérec in the 1990s.

MEDIA AND PUBLISHING

Television and radio. In 1989 the Socialist government formed the Supreme Audiovisual Council (Conseil Supérieur de l'Audiovisuel; CSA) to supervise radio and television broadcasting. There are both public and private stations. Programs also have been broadcast and received via satellite since 1984, and cable broadcasting began in 1987. More than three-fourths of the population watch television an average of 22 hours per week.

Although it has been largely eclipsed by television and video, radio still has cultural impact. Two agencies managed by Radio France—France Culture and France Musique—provide the bulk of the cultural programs, but they are often indifferently presented. Major stations such as France-Inter (public) or Europe No. 1 (private) have resorted increasingly to a mix of popular music, news items, quizzes, and talk shows. Smaller private stations cater to specialized interests—for example, Radio Notre Dame (religion) and Radio Classique (classical music). Popular music stations such as Fun Radio and Skyrock have grown rapidly. Since 1994, however, with the aim of protecting French culture, such stations have been obliged to dedicate 40 percent of their playlists to songs in French.

The press. The newspaper has a long history and a strong tradition in France. The French press, in the form of reviews and news sheets, has its origins in the early 17th century with Théophraste Renaudot's *La Gazette*, which began in 1631. It was not for another 250 years, however, with the passage of an act in 1881 allowing greater freedoms, that the press began to expand significantly. At the beginning of World War II, Paris offered some 30 daily papers, many with national followings and most with a clear political affiliation. The number of newspapers (and periodicals as well) declined sharply after the war, in some cases for political reasons but in others as a result of takeovers, collaborative ventures, and competition from television. In 1944 the Paris-based *Le Monde* was founded, and it became the most informed and influential of modern French newspapers. Other influential and widely circulating Paris dailies include *Le Figaro*, *Libération*, and *France-Soir*. Among the smaller dailies are the Roman Catholic *La Croix l'Événement* and the communist *L'Humanité*. In the 1950s illustrated magazines began to proliferate (echoing a trend of the 1930s); some of these were popular magazines of general interest and some were directed at specific markets, such as *Elle*, *Marie-Claire*, and *Vogue Paris* for women and *L'Express*, *Le Point*, and *Le Nouvel Observateur*, which are political. Few, however, have enjoyed the popular success and wide distribution of the news-oriented *Paris-Match*. By the late 20th century, three specific factors characterized the French press: first, the expansion of the regional daily paper, with *Ouest-France* enjoying the largest circulation in the country; second, the growth of specialized magazine journalism; and third, the appearance since the early 1960s of free newspapers essentially for advertising purposes, which are distributed weekly in the millions.

(Jo.E.F./J.N.T.)

For statistical data on the land and people of France, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Gaul

Geographic-historical scope. Gaul, in this context, signifies only what the Romans, from their perspective, termed Transalpine Gaul (Gallia Transalpina), or "Gaul across the Alps." Broadly, it comprised all lands from the Pyrenees and the Mediterranean coast of modern France to the English Channel and from the Atlantic to the Rhine and the western Alps. The Romans knew a second Gaul, Cisalpine Gaul (Gallia Cisalpina), or "Gaul this side of the Alps," in northern Italy—which, however, does not belong to the history of France. Transalpine Gaul came into existence as a distinct historical entity in the middle of the 1st century BC, through the campaigns of Julius Caesar (lived *c.* 100–44 BC), and disappeared late in the 5th century AD. Caesar's heir, the emperor Augustus (reigned 27 BC–AD 14), divided the country into four administrative provinces: Narbonensis, Lugdunensis, Aquitania, and Belgica. Following recognition of the impossibility of large-scale expansion beyond the Rhine, rulers of the Flavian dynasty (69–96) annexed the region between the middle Rhine and upper Danube, roughly the Black Forest region, to secure communications between Roman garrisons by then permanently established on both rivers. This area was called the Agri Decumates—Ten Cantons (the precise significance of the name being unknown). Its eastern border, conventionally referred to as the limes, assumed its final shape, as a defended palisade and ditch, under Antoninus Pius (138–161). The Agri Decumates were attached to Upper Germany (Germania Superior), one of two new frontier provinces (the other being Lower Germany [Germania Inferior]) created by the last Flavian emperor, Domitian (81–96). For greater administrative efficiency, the emperor Diocletian (284–305) subdivided all six Gallic provinces, forming a total of 13.

Agri Decumates

The people. Gaul was predominantly a Celtic land (see below), but it also contained pre-Celtic Ligurians and Iberians in the south and southwest and more recent Germanic immigrants in the northeast. Neighbouring Celtic communities on the Danube and in northern Italy, however, were not included. The south, in addition, had been heavily influenced by the Greek colony of Massilia (Marseille, founded around 600 BC) and its daughter-cities. In brief, the Gaul that was the foundation of medieval France was not a "natural" unit but a Roman construct, the result of a decision to defend Italy from across the Alps.

THE ROMAN CONQUEST

In the 2nd century BC Rome intervened on the side of Massilia in its struggle against the tribes of the hinterland, its main aim being the protection of the route from Italy to its new possessions in Spain. The result was the formation, in 121 BC, of "The Province" (Provincia, whence Provence), an area spanning from the Mediterranean to Lake Geneva, with its capital at Narbo (Narbonne). From 58 to 50 BC Caesar seized the remainder of Gaul. Although motivated by personal ambition, Caesar could justify his conquest by appealing to deep-seated Roman fear of Celtic war bands (see below) and further Germanic incursions (late in the 2nd century BC the Cimbri and Teutoni had invaded The Province and threatened Italy). Because of chronic internal rivalries, Gallic resistance was easily broken, though Vercingetorix' Great Rebellion of 52 had notable successes before it expired in the cruel siege of Alesia (Alise-Sainte-Reine).

Caesar's conquest of Gaul

Gaul under the High Empire (*c.* 50 BC–*c.* AD 250). The first centuries of Roman rule were remarkable for the speedy assimilation of Gaul into the Greco-Roman world. This was a consequence of both the light hand of the Roman imperial administration and the highly receptive nature of Gallic-Celtic society. Celtic culture had originated on the upper Danube around 1200 BC. Its expansion westward and southward, through diffusion and migration, was stimulated by a shift from bronze- to ironworking. Archaeologically, the type of developing Celtic Iron Age culture conventionally classified as Hallstatt appeared in Gaul from about 700 BC; in its La Tène form it made itself felt in Gaul after about 500 BC. Initially the Romans, who had not forgotten the capture of their city by Brennus, the leader of Celtic war bands, about 390 BC, despised and feared the Celts as barbarian savages. Until the end of the 1st century BC, they disparaged Gaul beyond The Province as Gallia Comata ("Long-Haired Gaul"), mocked and exploited the Gauls' craving for wine, and generally mismanaged The Province itself.

Gaul by then, however, was not far behind Rome in its evolution. In the south, Ligurian communities had long emulated the Hellenic culture of Massilia, as may be seen in the settlement of Entremont (near Aquae Sextiae [Aix-en-Provence]). In the Celtic core, Caesar found large nations (his *civitates*) coalescing out of smaller tribes (*pagi*) and establishing urban centres (*oppida—e.g.*, Bibracte [Mont Beuvray], near Augustodunum [Autun]), which, though quite unlike the classical city-states, were assuming significant economic and administrative functions. After the corrupt Roman Republic was replaced by the empire and its more prudent rule, these advances in Transalpine Gaul could be exploited for the imperial good. The Province, now Narbonensis, was planted with settlements of retired Roman soldiers (*coloniae*, "colonies"—

e.g., Arelate [Arles]); it soon became a land of city-states and was comparable with Italy in its way of life. In the remaining "Three Gauls," such colonies were few; there the *civitates* were retained, as was the habit of fierce rivalry between their leaders. Competition, however, was diverted from war: status was now measured in terms of the level of Romanization attained by both the individual and his community.

Civitas-capitals

Northern Gaul therefore became a Romanized land too. This is dramatically reflected in the dominance of Latin as the language of education and government; French was to be a Romance tongue. Archaeologically, however, Romanization in Gaul is most evident in the emergence of the Greco-Roman city. Although the *civitates* were too large to act as true city-states, they contained towns, either already in existence (*e.g.*, Lutetia Parisiorum [Paris]) or newly founded (*e.g.*, Augustodunum, "Augustusville"), that could be designated as their administrative centres and developed, by local magnates at their own expense, in accordance with classical criteria. Thus, these *civitas*-capitals, as scholars term them, were characterized by checkerboard street grids and imposing administrative and recreational buildings such as forums, baths, and amphitheatres. Although they display vernacular architectural traits, they essentially follow the best Mediterranean fashion. Most were unwalled—an indicator of the Pax Romana.

The mark of Rome is also discernible in the countryside, in the shape of villas. Villas of this period were, however, working farms as much as Romanized country residences—manor houses, not palaces. The survivors of the great Gallic aristocracy of the pre-Roman period, who first adopted Roman ways and who might eventually have constructed rural palaces, persisted into the 1st century AD but then seem to have been eclipsed by lesser landowners.

Scholars dispute the extent to which the mass of the Gallic population (at about 10 million, or 15 persons per square kilometre [39 persons per square mile], large for a preindustrial economy), free or slave, benefited from the new conditions, but there is no doubt that the landowners prospered. One of the great engines of their wealth was the Rhine army, which stimulated trade by purchasing its supplies from the interior. Commerce was greatly facilitated by a road network and system of river transport that had been expanded and improved under Roman administration. It is no accident that the capital of high imperial Gaul was Lugdunum (Lyon), the main Gallic road junction and a great inland port on the river route that led north to Colonia Agrippina (Cologne), the chief city of the two Germanies.

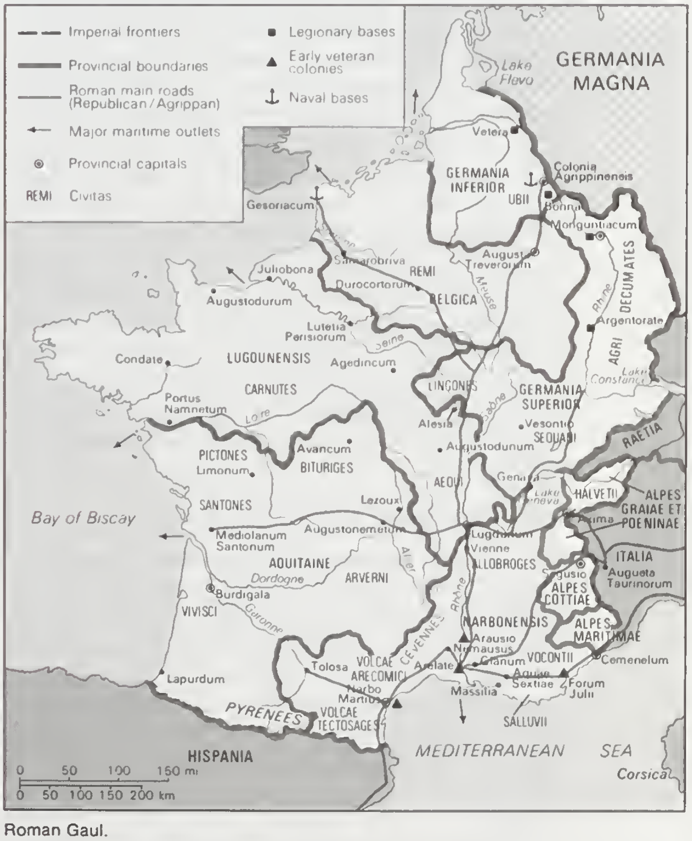
It is not surprising, therefore, that there was relatively little resistance to Roman rule; Vercingetorix was not an ancient Gallic hero. There were localized revolts in AD 21 and 69–70, but these were easily suppressed. They may have accelerated the demise of the old Gallic aristocracy; few Gauls subsequently pursued imperial Roman careers (for example, as senators). This diffidence, due initially perhaps to lingering Roman prejudice against Celts but reinforced by Gallic contentment with local responsibilities, may have served to keep Gallic wealth in Gaul.

Civil war and revolt

Gaul under the late Roman Empire (*c.* 250–*c.* 400). High Roman Gaul came to an end in an empirewide crisis characterized by foreign invasions and a rapid succession of rulers, as increased pressure on the empire's frontiers exacerbated its internal economic and political weaknesses. Priority was given to holding the Danube and the East; despite sporadic visits by emperors, the West was neglected. In 260 and 276 Gaul suffered depredation by two recent confederations of Germanic peoples, the Alemanni and the Franks (facing Upper and Lower Germany respectively). Civil war resulted in Gaul, Britain, and (for a while) Spain being governed by a line of "Gallic" emperors (beginning with Postumus [260–268]) before being reconquered by Aurelian in 274; there was further revolt in about 279–80. Although unity was reestablished and order of a sort restored by Aurelian (270–275), Probus (276–282), and Carinus (283–285), the country was much altered. For example, around 260 the Agri Decumates were abandoned, and, from about the reign of Probus, there began an extensive program of city fortification, though on very restricted circuits that cut through, and even used as building material, the proud structures of the previous age. The countryside was prey to marauding peasants. There was, however, no move to exploit the crisis to gain independence: the "Gallic Empire," though closely involving leading Gallic civilians, depended on the loyalty of the Rhine army; it thus championed Gallo-Roman, not Gallic, interests (essentially, the maintenance of a strong Rhine frontier).

After Diocletian and his successors radically reformed the empire in the late 3rd and early 4th centuries, Gaul enjoyed a new stability and even an enhanced role in imperial life. The reason for this was the empire's renewed commitment to defend Italy from the Rhine. To ensure the loyalty of the Rhine garrison and the civil population that depended on it for protection, imperial representation in the frontier region became permanent. An official of the highest rank, a praetorian prefect, was based there, and a series of emperors and usurpers (in particular, Constantine I [reigned 306–337], Julian [355–363], Valentinian I [364–375], Gratian [375–383], and Magnus Maximus [383–388]) resided there for at least part of their reigns. Their seat of government was usually Augusta Treverorum (Trier [German]; Trèves [French]), the former *civitas*-capital of the Treveri and capital of Belgica, now "the Rome of the West." (An interesting exception to the rule was Julian, who, with Trier rendered inhospitable by war, wintered in Paris, giving that city its first taste of future greatness.) Throughout the 4th century and especially in its latter half, the ever-present German menace as well as internecine strife occasionally caused the Rhine frontier to be broken, but it was always vigorously restored.

Some recovery of economic prosperity occurred, though it was fragile and uneven. The levying of taxes in kind rather than in cash may have weakened commerce, and the settlement of captive barbarians on the land indicates a rural labour shortage. Trier was endowed with magnificent buildings, but most Gallic cities never recovered their classical grandeur. The well-to-do, who were for the most part probably not descended from the aristocracy of high Roman Gaul (destroyed in the 3rd-century crisis), had



Roman Gaul.

Adapted from Westermann Grosser Atlas zur Weltgeschichte, Georg Westermann Verlag, Braunschweig loftier ambitions than their predecessors. Looking beyond the *civitates,* they eagerly sought posts in the imperial administration, now conveniently close to hand, basing their claim to advancement on their learning. (Gallo-Roman education, drawing vitality from the Gallo-Celtic love of eloquence, had long been renowned, but it blossomed fully in the 4th century in famous universities such as the one at Burdigala [Bordeaux].) As the century progressed, some educated Gauls grew extremely powerful; the best known, Ausonius (lived *c.* 310–*c.* 393), a poet and professor at Bordeaux, was appointed tutor of the future emperor Gratian and became his counselor. These worldly aristocrats, when not at court, favoured the country life; the later 4th century saw the rise of the palatial villa, especially in the southeast. Other Gauls looked to serve an even higher power; Christianity took root deeply in the land at this time. An episcopal hierarchy (based on the Roman provinces and *civitates*) was developed and monasticism was introduced by Martin of Tours (lived *c.* 316/336–397).

THE END OF GAUL (C. 400–C. 500)

From 395 the division of the Roman Empire between an eastern and a western half caused acute internal political stresses that encouraged barbarian penetration of the Danube region and even Italy. The Rhine frontier was again neglected, and the seat of the Gallic prefecture was moved to Arelate. The result was Germanic invasion and civil war. By 418, Franks and Burgundians were established west of the Rhine, and the Visigoths settled in Aquitania (Aquitaine). These Germans, however, were nominally allies of the empire, and, owing mainly to the energy of the Roman general Flavius Aetius, they were kept in check. The death of Aetius in 454 and the growing debility of a western imperial government hamstrung by the loss of Africa to the Vandals created a power vacuum in Gaul. It was filled by the Visigoths, at first indirectly through the nomination of the emperor Avitus (455–456) and then directly by their own kings, the most important being King Euric (466–484). The 460s and 470s saw steady Visigothic encroachment on Roman territory to the east; the Burgundians followed suit, expanding westward from Sapaudia (now Savoy). In 476 the last imperial possessions in Provence were formally ceded to the Visigoths.

Germanic invasions

Gaul suffered badly from these developments. Communities near the Rhine were destroyed by war. Refugees fled south, to Roman territory, only to find themselves burdened by crippling taxation and administrative corruption. As is evident from the works of the writer Sidonius Apollinaris (lived *c.* 430–*c.* 490), however, the economic power and with it the life-style of the Romano-Gallic aristocracy remained remarkably resilient, whether under Roman emperors or barbarian kings. Many aristocrats, as, for example, Sidonius himself, also confirmed their standing in their communities by becoming bishops. Until the middle of the 5th century, the leaders of Gallic society, lay and clerical, while learning to live with the barbarian newcomers, still looked to Rome for high office and protection. Thereafter they increasingly cooperated with the German rulers as generals and counselors. Thus, at least in the centre and south of the country, the Gallo-Roman cultural legacy was bequeathed intact to the successor-kingdoms. (J.F.Dr.)

Merovingian and Carolingian age

The period of the Merovingian and uninterrupted Carolingian Frankish dynasties (476–887) encompasses what scholars call the Early Middle Ages. After the 4th and 5th centuries, when Germanic peoples entered the Roman Empire in substantial numbers and brought the existence of that Mediterranean state to an end, the Franks played a key role in Gaul, unifying it under their rule. Merovingian and, later, Carolingian monarchs created a polity centred in an area between the Loire and Rhine but extending beyond the Rhine into large areas of Germany.

ORIGINS

Early Frankish period. In the second quarter of the 5th century, various groups of Franks moved southward toward the middle Rhine area (Cologne), the lower branches of the Moselle and Meuse, and the Atlantic coastal region. In the latter area, separate groups took possession of Tournai and Cambrai and reached the Somme. These Franks along the coast were divided into many small kingdoms. One of the better-known groups established itself in and around the *urbs* of Tournai; its leader (*regulus*) was Childeric (d. *c.* 481/482), who traditionally is regarded as a close relative in the male line of Merovech, eponymous ancestor of the Merovingian dynasty. Childeric placed himself in the service of the Roman Empire.

Ancestor of the Merovingian dynasty

Gaul and Germany at the end of the 5th century. Preceding the arrival of the Franks, other Germans had already entered Gaul. The area south of the Loire was divided between two groups. One, the Visigoths, occupied Aquitaine, Provence, and most of Spain. Their king, Euric (ruled 466–484), was the most powerful monarch in the West. The other group, the Burgundians, ruled much of the Rhône valley. In northern Gaul the Alemanni occupied Alsace and moved westward into the area between the Franks and Burgundians, while the first British immigrants established themselves on the Armorican Peninsula (now Brittany). Substantial parts of Gaul were ruled by Syagrius, a Roman king (*rex*), with his capital at Soissons.

In spite of the German influx, Gaul, which had been part of the Roman Empire for about 500 years, remained thoroughly Romanized. Because many of its administrative institutions withstood the crisis of the 5th century, Gaul's traditional Roman civilization survived, at least in attenuated form, especially among the aristocratic classes. The core of political, social, economic, and religious life remained in the *civitas* with the *urbs* at its heart. In addition, the Germans themselves were, to varying degrees, affected by this civilization. This influence was stronger among the Burgundians and the Visigoths, who had lived within the empire for a longer time and had intermingled with other Germanic peoples to a great extent, than it was among the Franks and Alemanni, who had only recently entered the empire. The Burgundians and Visigoths adopted a heretical form of Christianity—Arianism. The Franks and Alemanni remained pagan and preserved limited contacts with Germans living outside the boundaries of the Roman Empire.

Survival of Roman civilization

In effect, the Germanic peoples who penetrated into Roman Gaul were but a small segment of the Germanic world. The northern Germans (Angles, Jutes, Saxons, and Frisians) still occupied the coastal regions of the North Sea east of the Rhine; the Thuringians and Bavarians divided the territory between the Elbe and Danube; the Slavic world began on the opposite bank of the Elbe.

THE MEROVINGIANS

Clovis. Clovis (ruled 481/482–511), the son of Childeric, unified Gaul with the exception of areas in the southeast.

Frankish expansion and the unification of Gaul. During the years following his accession, Clovis consolidated the position of the Franks in northern Gaul. In 486 he defeated Syagrius, the last Roman ruler in Gaul, and in a series of subsequent campaigns with strong Gallo-Roman support occupied an area situated between the Frankish kingdom of Tournai, the Visigothic and Burgundian kingdoms, and the lands occupied by the Rhenish (Ripuarian) Franks and the Alemanni, removing it from imperial control once more. It was probably during this same period that he eliminated the other Salian kings. In a second phase he attacked the other Germanic peoples living in Gaul, with varying degrees of success. An Alemannian westward push was blocked, probably as a result of two campaigns—one conducted by the Franks of the kingdom of Cologne in about 495–496 at the Battle of Tolbiacum (Zülpich), the second by Clovis in about 506, after his annexation of Cologne. Clovis thus extended his authority over most of the territory of the Alemanni. Some of the former inhabitants sought refuge in Theodoric's Ostrogothic kingdom. After his conversion to Roman Christianity about 498, Clovis absorbed the region between the Seine and the Loire (including Nantes, Rennes, and Vannes) and then moved against the Visigothic kingdom. He defeated Alaric II at Vouillé (507). He annexed Aquitaine, between the Loire, Rhône, and Garonne, as well as Novempopulana, between the Garonne and the Pyrenees. Opposed to a Frankish hegemony in the West, Theodoric intervened on behalf of the Visigothic king. He prevented Clovis from annexing Septimania, on the Mediterranean between the Rhône and the Pyrenees, which the Visigoths retained, and occupied Provence. In addition, Clovis eliminated various Frankish kinglets in the east and united the Frankish people under his own leadership.

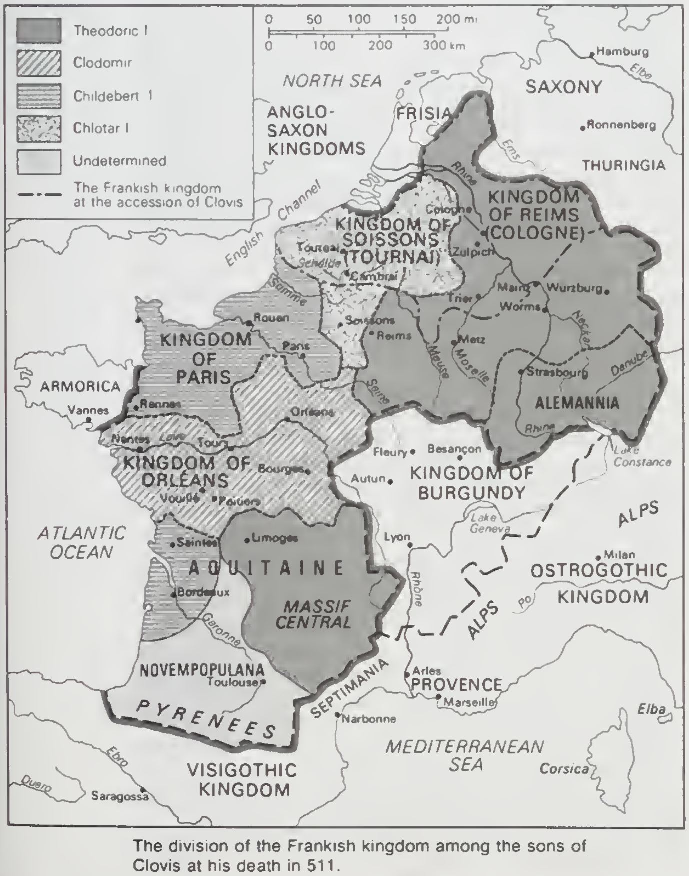
Clovis established Paris as the capital of his new kingdom, and about 507–508 he received from the emperor an honorary consulship and the right to use the imperial insignia. These privileges gave the new king legitimacy of sorts and were useful in gaining the support of his Gallo-Roman subjects.

Triumph of Roman Catholicism in Gaul

The conversion of Clovis. Clovis came to believe that his victory at Tolbiacum in 496 was due to the help of the Christian God, whom his wife Clotilda had been encouraging him to accept. With the support of Bishop Remigius of Reims, a leader of the Gallo-Roman aristocracy, Clovis converted with some 3,000 of his army. Clovis' conversion assured the Frankish king of the support not only of the ecclesiastical hierarchy but in general also of Roman Christians—the majority of the population. It also ensured the triumph in Gaul of Roman Christianity over paganism and Arianism and spared Gaul the lengthy conflicts that occurred in other Germanic kingdoms.

The sons of Clovis. Following the death of Clovis in 511, the kingdom was divided among his four sons. This partition was not made according to ethnic, geographic, or administrative divisions. The only factor taken into account was that the portions be of equal value (defined in terms of the royal fisc, which had previously been the imperial fisc, and tax revenues from land and trade, which were based upon imperial practices). Boundaries were poorly defined. The territory was divided into two general areas: one was the territory north of the Loire (the part of Gaul that was conquered earliest); the other, to the south in Aquitaine, was a region not yet assimilated. Theodoric I, Clovis' eldest son by one of his wives married in Germanic style before Clovis married Clotilda and converted to Christianity, received lands around the Rhine, Moselle, and upper Meuse, as well as the Massif Central; Clodomir, the Loire country to the other side of the Rhine (this kingdom was the only one not composed of separated territories); Childebert I, the country of the English Channel and the lower Seine and, probably, the region of Bordeaux and Saintes; Chlotar I, the old Frankish country north of the Somme and an ill-defined area in Aquitaine. Their capitals were centred in the Paris Basin, which was divided among the four brothers: Theodoric I used Reims; Clodomir, Orléans; Childebert I, Paris; Chlotar I, Soissons. As each brother died, the survivors partitioned the newly available lands among themselves. This system resulted in bloody competition until 558, when Chlotar I, after his brothers' deaths, succeeded in reuniting the kingdom under his own rule.

From R Grousset and E Leonard (eds), *Histoire Universelle*



The division of the Frankish kingdom among the sons of Clovis at his death in 511.

The conquest of Burgundy and southern Germany. In spite of these partitions, the Frankish kings continued their conquests. One of their primary concerns was to extend their dominion over the whole of Gaul. It took two campaigns to overcome the Burgundian kingdom. In 523 Clodomir, Childebert I, and Chlotar I, as allies of Theodoric, king of the Ostrogoths, moved into Burgundy, whose king, Sigismund, Theodoric's son-in-law, had assassinated his own son. Sigismund was captured and killed. Godomer, the new Burgundian king, defeated the Franks at Vézeronce and forced them to retreat; Clodomir was killed in the battle. Childebert I, Chlotar I, and Theodebert I, the son of Theodoric I, regained the offensive in 532–534. The Burgundian kingdom was annexed and divided between the Frankish kings. Following Theodoric's death in 526, the Franks were able to gain a foothold in Provence by taking advantage of the weakened Ostrogothic kingdom. The Franks were thus masters of all southeastern Gaul and had reached the Mediterranean. But, in spite of two expeditions (531 and 542: the siege of Saragossa), they were unable to gain possession of Visigothic Septimania. Also, at least a portion of Armorica in the northwest remained outside the Frankish sphere of influence. During this period, British colonization of the western half of the Armorican Peninsula was at its height.

Conquests of Thuringia and Bavaria

To the east, the Franks extended their domain in southern Germany, bringing Thuringia (*c.* 531 Chlotar I carried off Radegunda, a niece of the Thuringian king), the part of Alemannia between the Neckar and the upper Danube (after 536), and Bavaria under subjection. The latter was created as a dependent duchy in about 555. The Franks were less successful in northern Germany; in 536 they imposed a tribute on the Saxons (who occupied the area between the Elbe, the North Sea, and the Ems), but the latter revolted successfully in 555.

Theodebert I and his son, Theodebald, sent expeditions into Italy during a struggle between the Ostrogoths and Byzantines (535–554), but they achieved no lasting results.

The grandsons of Clovis. At the death of Chlotar I (561) the Frankish kingdom, which had become the most powerful state in the west, was once again divided, this time between his four sons. The partition agreement, based on that of 511, dealt with more extensive territories. Guntram received the eastern part of the former kingdom of Orléans, enlarged by the addition of Burgundy. Charibert I's share was fashioned from the old kingdom of Paris (Seine and English Channel districts), augmented in the south by the western section of the old kingdom of Orléans (lower Loire valley) and the Aquitaine Basin. Sigebert I received the kingdom of Reims, extended to include the new German conquests; a portion of the Massif Central (Auvergne) and the Provençal territory (Marseille) were added to his share. Chilperic I's portion was reduced to the kingdom of Soissons.

The death of Charibert (567) resulted in still a further partition. Chilperic, the principal beneficiary, received the lower Seine district, including a large tract of the English Channel coast. The remainder, most notably Aquitaine and the area around Bayeux, was divided in a complex manner; and Paris was subject to joint possession. The partitions of 561 and 567, which reaffirmed the division of Francia, were the sources of innumerable intrigues and family struggles, especially between, on the one hand, Intrigues and family struggles

Chilperic I, his wife Fredegund, and their children, who controlled northwestern Francia, and, on the other hand, Sigebert I, his wife Brunhild, and their descendants, the masters of northeastern Francia.

The shrinking of the frontiers and peripheral areas. These events undermined the Frankish hegemony. In Brittany the Franks maintained control of the eastern region but had to cope with raids by the Bretons, who had established heavily populated settlements in the western part of the peninsula. To the southwest the Gascons, a highland people from the Pyrenees, had been driven northward by the Visigoths in 578 and settled in Novempopulana; in spite of several Frankish expeditions, this area was not subdued. In the south the Franks were unable to gain control of Septimania; they tried to accomplish this by means of diplomatic agreements, which were buttressed by dynastic intermarriage, and by military campaigns occasioned by religious differences (the Visigothic kings were Arians). In the southeast the Lombards, who had recently arrived in Italy, made several raids on Gaul (569, 571, 574); Frankish expeditions into Italy (584, 585, 588, 590), led by Childebert II, were without result. Meanwhile the Avars, a people of undetermined origin who settled along the Danube in the second half of the 6th century, threatened the eastern frontier; in 568 they took Sigebert prisoner, and in 596 they attacked Thuringia, forcing Brunhild to purchase their departure.

The parceling up of the kingdom. Internal struggles resulted in the emergence of new political configurations. At the time of the partitions of 561 and 567, new political geographic units began to appear within Gaul. Austrasia was created from the Rhine, Moselle, and Meuse districts, which had formerly been the kingdom of Reims, and the areas east of the Rhône conquered by Theodoric and his son Theodebert; Sigebert I (d. 575) transferred its capital to Metz in order to take advantage of the income provided by trade on the Rhine. Neustria was born out of the partition of the kingdom of Soissons; a portion of the kingdom of Paris was added to it, thus endowing the area with a broad coastal section and making the lower Seine valley its centre. Its first capital, Soissons, was returned to Austrasia following the death of Chilperic I; its capital was later moved to Paris, which had been controlled by Chilperic. The kingdom of Orléans, without its western territory but with part of the old Burgundian lands added to it, eventually became Burgundy; Guntram fixed its capital at Chalon-sur-Saône. Aquitaine submitted to the Frankish kingdoms centred farther north in Gaul; its *civitates* were the object of numerous partitions made by sovereigns who regarded it as an area for exploitation. Aquitaine did not enjoy political autonomy during this period.

The failure of reunification (613–714). Territorial crisis was partially and provisionally averted during the first third of the 7th century.

Chlotar II and Dagobert I. Chlotar II, king of Neustria since 584, took control of Burgundy and Austrasia in 613 upon the brutal execution of Brunhild, and thus a united kingdom once again was created. He fixed his capital at Paris and, in 614, convoked a council there, at which he recognized the traditional prerogatives of the aristocracy (Gallo-Roman and Germanic) in order to gain their support in the governing of the kingdom. His son Dagobert I (ruled 629–639) was able to preserve this unity; he journeyed to Burgundy, where the highest political office, mayor of the palace, was maintained, to Austrasia, and to Aquitaine, which was given the status of a duchy. He thus recognized structures of imperial origin.

Dagobert had only limited success along the frontier. In 638 he placed the Bretons and the Gascons under nominal subjection, but ties with these peripheral regions were tenuous. He intervened in dynastic quarrels of Spain, entering the country and going as far as Saragossa before receiving tribute and quitting the country. Septimania remained Visigothic. On the eastern frontier there were incidents involving Frankish merchants and Moravian and Czech Slavs; aftcr the failure of a campaign conducted by Dagobert, with the assistance of the Lombards and Bavarians (633), the Slavs attacked Thuringia. The king reached an agreement with the Saxons, who would protect the eastern frontier in return for remission of a tribute they had paid since 536. Thus Dagobert used traditional imperial techniques to protect the frontiers with more or less Romanized barbarians.

The hegemony of Neustria. The territorial struggles began anew after 639. In Neustria, Austrasia, and Burgundy, power was in the hands of aristocratic leaders, called the mayors of the palace. Ebroïn, mayor of the palace in Neustria, attempted to unify the kingdom under his leadership but met with violent opposition. Resistance in Burgundy was led by Bishop Leodegar, who was assassinated in about 679. (He was later canonized.) Austrasia was governed by the Pepinid mayors of the palace; Pepin I of Landen was succeeded by his son Grimoald, who tried unsuccessfully to have his son, Childebert the Adopted, crowned king, and by Pepin II of Herstal (or Héristal), whom Ebroïn was briefly able to keep from power (*c.* 680).

Struggle between Neustria and Austrasia

Frankish hegemony was once more threatened in the peripheral areas, especially to the east where Austrasia was endangered. The Thuringians (640–641) and Alemanni regained their independence. The Frisians reached the mouth of the Schelde and controlled the towns of Utrecht and Dorestat; the attempted conversion of Frisia by Wilfrid of Northumbria had to be abandoned (*c.* 680). In southern Gaul, Duke Lupus changed the status of Aquitaine from that of a duchy to an independent principality.

Austrasian hegemony and the rise of the Pepinids. The murder of Ebroïn (680 or 683) reversed the situation in favour of Austrasia and the Pepinids. Pepin II, who defeated the Neustrians at Tertry in 687, reunified northern Francia under his own control during the next decade. Austrasia and Neustria were reunited under a Merovingian figurehead, but Pepin II governed by virtue of his position as mayor of the palace. At the same time, Pepin II partially restabilized the frontiers of northern Francia by driving the Frisians north of the Rhine and by restoring Frankish suzerainty over the Alemanni. But control of southern Gaul continued to elude Pepin II and his supporters. In the early 8th century, Provence became an autonomous duchy, while power in Burgundy was divided.

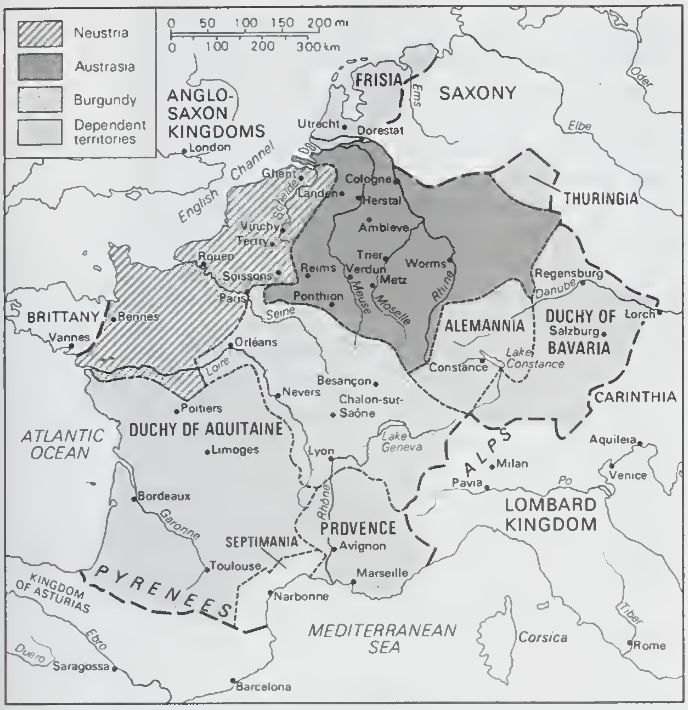
Pepin II

THE CAROLINGIANS

Representatives of the Merovingian dynasty continued to hold the royal title during most of the first half of the 8th century. They were captive monarchs, according to a contemporary biographer of Charlemagne. In actual fact, effective power was in the hands of the Pepinids, who, thanks to their valuable landholdings and loyal retainers, maintained a monopoly on the office of mayor of the palace. Because of their familial predisposition for the name Charles and because of the significance of Charlemagne in the family's history, modern historians have called them the Carolingian dynasty.

Charles Martel and Pepin III the Short. Pepin II's death in 714 jeopardized Carolingian hegemony. His heir was a grandchild entrusted to the regency of his widow, Plectrude. There was a revolt in Neustria, and Eudes, duke of Aquitaine, used the occasion to increase his holdings and make an alliance with the Neustrians. The Saxons crossed the Rhine and the Arabs crossed the Pyrenees.

Charles Martel. The situation was rectified by Pepin's illegitimate son, Charles Martel. Defeating the Neustrians at Amblève (716), Vincy (717), and Soissons (719), he made himself master of northern Francia. He then reestablished Frankish authority in southern Gaul, where the local authorities could not cope with the Islāmic threat; he stopped the Muslims near Poitiers (Battle of Tours; 732) and used this opportunity to subdue Aquitaine (735–736). The Muslims then turned toward Provence, and Charles Martel sent several expeditions against them. At the same time, he succeeded in reestablishing authority over the dissident provinces in the southeast (737–738) with the exception of Septimania. Finally, he reestablished his influence in Germany. In his numerous military campaigns he succeeded in driving the Saxons across the Rhine, returned the Bavarians to Frankish suzerainty, and annexed southern Frisia and Alemannia. He also encouraged missionary activity, seeing it as a means to consolidate his power; this undertaking was supported by the papacy,



The Frankish domains in the time of Charles Martel (boundaries approximate).

From G. Fournier, *L'Occident de la fin du V^e siecle a la fin du IX^e siecle*

which was beginning to seek support in the West. Missionaries east of the Rhine, most of whom were Anglo-Saxon (*e.g.,* Willibrord and Winfrid, also known as Boniface), made definite progress in their task.

Charles Martel's chief source of power

Charles Martel had supported a figurehead Merovingian king, Theodoric IV (ruled 721–737), but upon the latter's death he felt his own position secure enough to leave the throne vacant. His chief source of power was a strong circle of followers, who furnished the main body of his troops and became the most important element in the army because local dislocation of government had weakened the recruitment of the traditional levies of free men. He attached them to himself by concessions of land, which he obtained by drawing on the considerable holdings of the church. This gave him large tracts of land at his disposal, which he granted for life (*precaria*). He was thus able to recruit a larger and more powerful circle of followers than that surrounding any of the other influential magnates.

Pepin III the Short. At the death of Charles Martel (741), as was the custom, the lands and powers in his hands were divided between his two sons, Carloman and Pepin III the Short. This partition was followed by unsuccessful insurrections in the peripheral duchies—Aquitaine, Alemannia, and Bavaria.

Carloman's entrance into a monastery in 747 reunited Carolingian holdings. Pepin the Short, who had held de facto power over Francia, or the *regnum Francorum,* as mayor of the palace, now desired to be king. He was crowned with the support of the papacy, which, threatened by the Lombards and having problems with Byzantium, sought a protector in the West. The change of dynasty was accomplished in two stages: in 751, after obtaining the support of Pope Zacharias, Pepin deposed Childeric III; he then had himself elected king by an assembly of magnates and consecrated by the bishops, thus ending the nominal authority of the last Merovingian king, Childeric III, who had been placed on the throne in 743. The new pope, Stephen II (or III), sought aid from Francia; in 754 at Ponthion he gave Pepin the title patrician of the Romans, renewed the king's consecration, and consecrated Pepin's sons, thus providing generational legitimacy for the line.

As king, Pepin limited himself to consolidating royal control in Gaul, thus establishing the base for later Carolingian expansion. Despite Pepin's efforts, the situation at the German frontier was unstable. The duchy of Bavaria, which had been given to Tassilo III as a benefice, gained its independence in 763; several expeditions were unable to subdue the Saxons. On the other hand, Pepin achieved a decisive victory in southern Gaul by capturing Septimania from the Muslims (752–759). He broke down Aquitaine's resistance, and it was reincorporated into the kingdom (760–768). Pepin intervened in Italy twice (754–755; 756) on the appeal of the pope and laid the foundations for the Papal States. He exchanged ambassadors with the great powers of the eastern Mediterranean—the Byzantine Empire and the caliphate of Baghdad.

Charlemagne. Pepin III, faithful to ancient customs, divided his kingdom between his two sons, Charles (Charlemagne) and Carloman. On Carloman's death in 771 the kingdom was reunited. Charlemagne established the base of his kingdom in northeastern Francia (his preferred residence was Aachen [Aix-la-Chapelle]).

The conquests. Charlemagne extended considerably the territory he controlled and unified a large part of the Christian West; he followed no grand strategy of expansion, taking advantage, instead, of situations as they arose.

Territorial consolidation and conquests

Charlemagne consolidated his authority up to the geographic limits of Gaul. Though he put down a new insurrection in Aquitaine (769), he was unable to bring the Gascons and the Bretons fully under submission. He pursued an active policy toward the Mediterranean world. In Spain he attempted to take advantage of the emir of Córdoba's difficulties; he was unsuccessful in western Spain, but in the east he was able to establish a march south of the Pyrenees to the important city Barcelona. Pursuing Pepin's Italian policy, he intervened in Italy. At the request of Pope Adrian I, whose territories had been threatened by the Lombards, he took possession of their capital city, Pavia, and had himself crowned king of the Lombards. In 774 he fulfilled Pepin's promise and created a papal state; the situation on the peninsula remained unsettled, and many expeditions were necessary. This enlargement of his Mediterranean holdings led Charlemagne to establish a protectorate over the Balearic Islands in the western Mediterranean (798–799).

Charlemagne conquered more German territory and secured the eastern frontier. By means of military campaigns and missionary activities he brought Saxony and northern Frisia under control; the Saxons, led by Widukind, offered a protracted resistance (772–804), and Charlemagne either destroyed or forcibly deported a large part of the population. To the south, Bavaria was brought under Frankish authority and annexed. Conquests in the east brought the Carolingians into contact with new peoples—Charles was able to defeat the Avars in three campaigns (791, 795, 796), from which he obtained considerable booty; he was also able to establish a march on the middle Danube, and the Carolingians undertook the conversion and colonization of that area. Charles established the Elbe as a frontier against the northern Slavs. The Danes constructed a great fortification, the Dannevirke, across the peninsula to stop Carolingian expansion. Charles also founded Hamburg on the banks of the Elbe. These actions gave the Franks a broad face on the North Sea.

The Frankish state was now the principal power in the West. Charlemagne claimed to be defender of Roman Christianity and intervened in the religious affairs of Spain. Problems arose over doctrinal matters that, along with questions concerning the Italian border and the use of the imperial title, brought him into conflict with the Byzantine Empire; a peace treaty was signed in 810–812. Charles continued his peace policy toward the Muslim East: ambassadors were exchanged with the caliph of Baghdad, and Charles received a kind of eminent right in Jerusalem.

Charlemagne's coronation as emperor

The restoration of the empire. When by the end of the 8th century Charles was master of a great part of the West, he reestablished the empire in his own name. He was crowned emperor in Rome (Christmas Day, 800), by Pope Leo III. Charlemagne's powers in Rome and in relation to the Papal States, which were incorporated, with some degree of autonomy, into the Frankish empire, were clarified. Although his new title did not replace his royal titles, it was well suited to his preponderant position in the old Roman West. The imperial title indicates a will to unify the West; nevertheless, Charlemagne preserved the kingdom of Italy, giving the crown to one of his sons, Pepin, and made Aquitaine a kingdom for his other son, Louis. Emperors ruled over kings.

Louis I the Pious. Only chance ensured that the empire remained united under Louis I the Pious, the last surviving son of Charlemagne (the latter had anticipated the partitioning of his empire among his sons). The era of great conquests had ended, and, on the face of it, Louis's principal preoccupation was his relations with the peoples to the north. In the hope of averting the threat posed by the Vikings, who had begun to raid the coasts of the North Sea and the Atlantic Ocean, Louis proposed to evangelize the Scandinavian world. This mission was given to St. Ansgar but was a failure.

Louis's conception of the empire

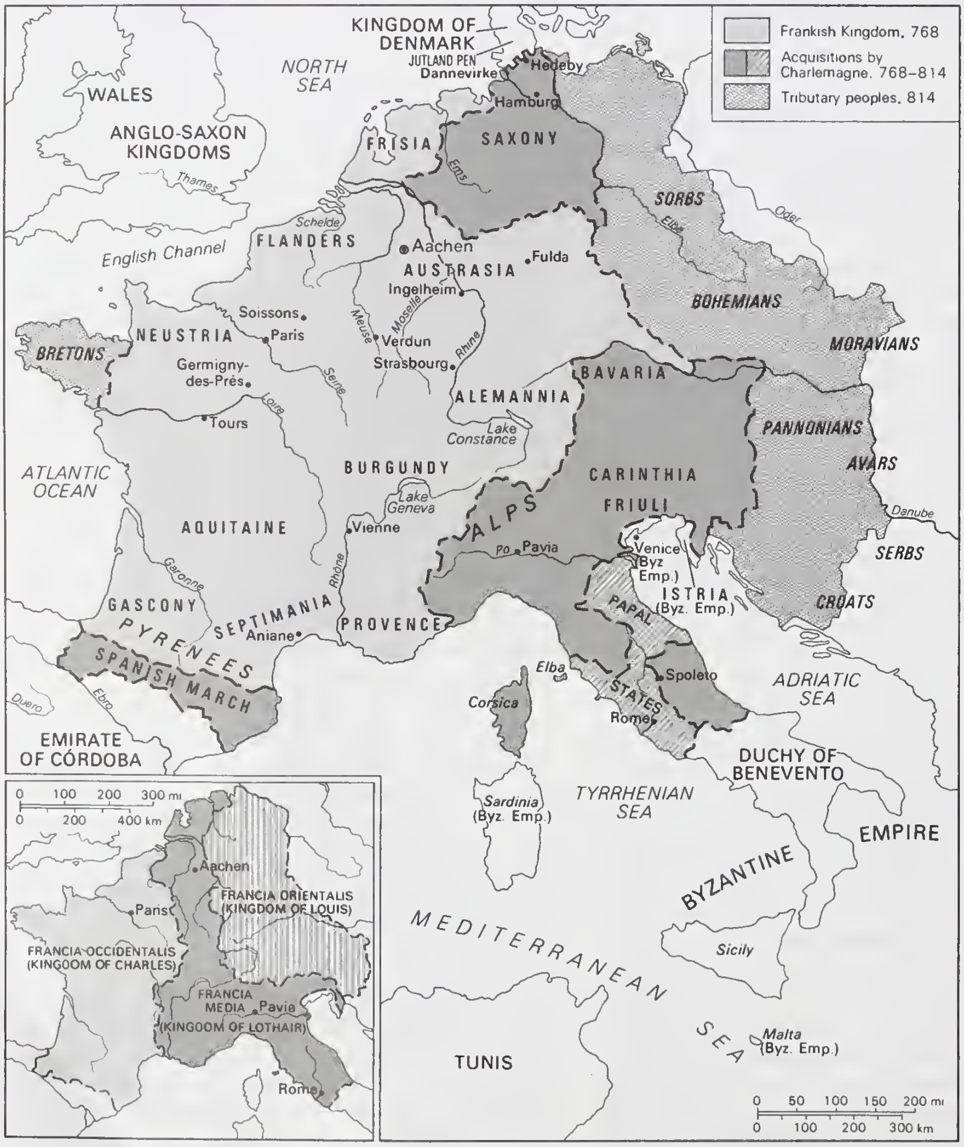
During Louis's reign, the imperial bureaucracy was given great uniformity. Louis saw the empire, above all, as a religious ideal, and in 816 the imperial coronation, originally a secular ceremony, was complemented by a religious ceremony, the anointment, at which the pope presided. At the same time Louis the Pious took steps to regulate the succession so as to maintain the unity of the empire (*Ordinatio Imperii,* 817). His oldest son, Lothair I, was to be sole heir to the empire, but within it three dependent kingdoms were maintained: Louis's younger sons, Pepin and Louis, received Aquitaine and Bavaria, respectively; his nephew Bernard was given Italy.

The remarriage of Louis the Pious to Judith of Bavaria and the birth of a fourth son, Charles II the Bald, upset this project. In spite of opposition from Lothair, who had the support of a unity faction drawn from the ranks of the clergy, the emperor's principal concern was to create a kingdom for Charles the Bald. These divergent interests led to conflicts that weakened imperial prestige (in 833, abandoned by his followers at the Field of Lies, Louis the Pious was forced to make public penance at the church of Notre-Dame at Compiègne). The question of Aquitaine arose at the death of Pepin I, ruler there since 814; the emperor gave this subordinate kingdom to Charles, but the magnates rose up and proclaimed Pepin II, the son of the dead king.

The partitioning of the Carolingian empire. After the death of Louis I the Pious (840), his sons continued their plotting to alter the succession. Louis II the German and Charles II the Bald affirmed their alliance against Lothair I (Oath of Strasbourg, 842).

The Treaty of Verdun. Later the three brothers came to an agreement in the Treaty of Verdun (843). The empire was divided into three kingdoms arranged along a north–south axis: Francia Orientalis was given to Louis, Francia Media to Lothair, and Francia Occidentalis to Charles the Bald. The three kings were equal among themselves. Lothair kept the imperial title, but it had completely lost its universal character and had meaning only in a portion of the old empire.

Adapted from p. 104, *The Middle Ages, 395–1500,* 5th ed., J.R. Strayer and D.C. Munro, copyright © 1970; by permission of Appleton-Century-Crofts, Educational Division, Meredith Corporation



The Carolingian empire and (inset) divisions after the Treaty of Verdun, 843.

The kingdoms created at Verdun. Until 861 the clerical faction tried to impose a government of fraternity on the descendants Charlemagne, manifested in the numerous conferences they held; but particularistic forces destroyed it.

Francia Media proved to be the least stable of the kingdoms, and the imperial institutions bound to it suffered as a result. In 855 the death of Lothair I was followed by a partition of his kingdom among his three sons: the territory to the north and west of the Alps went to Lothair II (Lotharingia) and to Charles (kingdom of Provence); Louis II received Italy and the imperial title. At the death of Charles of Provence (863), his kingdom was divided between his brothers Lothair II (Rhône region) and Louis II the German (Provence). After the death of Lothair II in 869, Lotharingia was divided between his two uncles, Louis the German and Charles the Bald. Louis, however, did not gain control of his share until 870. Charles was made master of the Rhône regions of the ancient kingdom of Provence. Louis II (d. 875) devoted most of his attention to fighting the Muslims who threatened the peninsula and the papal territories.

Viking invasions

In Francia Occidentalis Charles II the Bald was occupied with the struggle against the Vikings, who ravaged the countryside along the Scheldt, Seine, and Loire rivers. More often than not, the king was forced to pay for their departure with silver and gold. Aquitaine remained a centre of dissension. For some time (until 864) Pepin II continued to have supporters there, and Charles the Bald attempted to pacify them by installing his sons—first Charles the Young (ruled 855–866) and then Louis II the Stammerer (ruled 867–877)—on the throne of Aquitaine. The problems in Aquitaine were closely connected to general unrest among the magnates, who wished to keep the regional king under their control. By accumulating countships and creating dynasties, the magnates succeeded in carving out large principalities at the still unstable borders: Robert the Strong and Hugh the Abbot in the west; Eudes, son of Robert the Strong, in this same region and in the area around Paris; Hunfred, Vulgrin, Bernard Plantevelue, count of Auvergne, and Bernard of Gothia in Aquitaine and the border regions; Boso in the southeast; and Baldwin I in Flanders. Nevertheless, Charles the Bald appeared to be the most powerful sovereign in the West, and in 875 Pope John VIII arranged for him to accept the imperial crown. An expedition he organized in Italy on the appeal of the pope failed, and the magnates of Francia Occidentalis rose up. Charles the Bald died on the return trip (877). Charles's son, Louis II the Stammerer, ruled for only two years. At his death in 879 the kingdom was divided between his sons Louis III and Carloman. In the southeast, Boso, the count of Vienne, appropriated the royal title to the kingdom of Provence. The imperial throne remained vacant. The death of Louis III (882) permitted the reunification of Francia Occidentalis (except for the kingdom of Provence) under Carloman.

In Francia Orientalis royal control over the aristocracy was maintained. But decentralizing forces, closely bound to regional interests, made themselves felt in the form of revolts led by the sons of Louis the German. The latter had made arrangements to partition his kingdom in 864, with Bavaria and the East Mark to go to Carloman, Saxony and Franconia to Louis the Younger, and Alemannia (Swabia) to Charles III the Fat. Although Louis II the German managed to gain a portion of Lotharingia in 870, he was unable to prevent Charles the Bald's coronation as emperor (875).

When Louis the German died in 876, the partition of his kingdom was confirmed. At the death of Charles the Bald, Louis the German's son Carloman seized Italy and intended to take the imperial title, but ill health forced him to abandon his plans. His youngest brother, Charles III the Fat, benefited from the circumstances and restored the territorial unity of the empire. The deaths of his brothers Carloman (880) and Louis III the Younger (882) without heirs allowed him to acquire successively the crown of Italy (880) and the imperial title (881) and to unite Francia Orientalis (882) under his own rule. Finally, at the death of Carloman, son of Louis the Stammerer, Charles the Fat was elected king of Francia Occidentalis (885); the magnates had bypassed the last heir of Louis the Stammerer, Charles III the Simple, in his favour.

Charles the Fat avoided involving himself in Italy, in spite of appeals from the pope, and concentrated his attention on coordinating resistance to the Vikings, who had resumed the offensive in the valleys of the Scheldt, Meuse, Rhine, and Seine. He was unsuccessful, however, and in 886 had to purchase the Vikings' departure: they had besieged Paris, which was defended by Count Eudes. The magnates of Francia Orientalis rose up and deposed Charles the Fat in 887.

THE FRANKISH WORLD

Society. *Germans and Gallo-Romans.* The settlement of Germanic peoples in Roman Gaul brought people from two entirely different backgrounds into contact. Linguistic barriers were quickly overcome, for the Germans adopted Latin. At the same time, German names were preponderant. Although there were religious difficulties in those regions settled by peoples converted to Arianism (Visigoths, Burgundians), Clovis' conversion simplified matters. The Germans who settled in Gaul were able to preserve some of their own judicial institutions, but these were heavily influenced by vulgar Roman law. The first sovereigns committed the now Roman-influenced customs of the people to writing, in Latin (Code of Euric, *c.* 470–480; Salic Law of Clovis, *c.* 507–511; Law of Gundobad, *c.* 501–515) and occasionally had summaries of Roman rights drawn up for the Gallo-Roman population (Papian Code of Gundobad; Breviary of Alaric); this system of personality of the law, certain aspects of which survived well into the 9th century, was gradually replaced by a territorially based legal system. Multiple contacts in daily life produced an original civilization composed of a variety of elements, some of which were inherited from antiquity, some brought by the Germans, and many strongly influenced by Christianity.

The mixing of cultures

Social classes. The collapse of Roman imperial power and the influx of Germans did not destroy the old Roman senatorial and landed aristocracy; the 6th-century kings called on its members to serve in the administration. A sort of military aristocracy had existed among the Germans: at the time of their settlement within the empire, its members were given tax revenues and lands confiscated from the Gallo-Roman aristocracy or awarded from the fisc (royal treasury). The two groups fused rapidly. They shared a common life, discharging public and religious duties and frequenting the court. By the beginning of the 7th century there arose an aristocracy of office, whose signs of prestige were the possession of land and service to the king and church. This aristocracy increased in importance during the conflicts between the Merovingian sovereigns. The ascendance of the Pepinids, Carolingian rule, and the power struggles in the 9th century furnished these magnates, on whom those in power were dependent, with a means of enriching themselves and augmenting their political and social influence.

The aristocracy

Parallel to this class of lay magnates and largely drawn from the same families was an ecclesiastical aristocracy, which was one both of office and of land. The church found itself in possession of a vast landed fortune. At the beginning of the 7th century, at least, the church frequently benefited from immunity, and governmental rights were conferred on abbots or bishops.

A class of small and middle-sized landholders apparently existed, about which little is known. It appears that both the power of the magnates and the practices born of the ancient patronage system, combined with extensive military service, had the effect of diminishing the size of this class.

During the Merovingian epoch, slavery, inherited from antiquity, was still a viable institution. Slaves continued to be obtained in war and through trade. But the number of slaves decreased under the influence of the church, which encouraged manumission and sought to prohibit the enslavement of Christians. Under the Carolingians, the slaves in Gaul formed only a residual class, although the slave trade was still active. Taken increasingly from the Slavic territories (the term *slavus* replaced the traditional *servus*), slaves were a commodity for trade with the Muslim lands of the Mediterranean.

Vassalage

Diffusion of political power. During the period of insecurity and turbulence that marked the end of the Merovingian epoch, bonds of personal dependence, present in both Roman and Germanic institutions, can be seen to compete with weakened governmental institutions. In the 7th century these bonds took one of two forms: commendation (a free man placed himself under the protection of a more powerful lord for the duration of his life) and precarious contract (a powerful lord received certain services in return for the use of his land for a limited time under advantageous conditions). In the 8th century the Pepinids increased their personal circle of followers. Charlemagne attempted to encompass the entire free population of his empire within a vassalic relationship. He encouraged an increase in the number of royal vassals and gave them administrative functions. During the 9th-century power struggles, however, some administrative offices became hereditary, though this represented a distortion of the vassalic relationship. In addition, before the end of the century a man could place himself in vassalage to several lords. Finally, the usurpation of governmental powers led to the formation of territorial principalities, resulting in a great weakening of royal authority.

Institutions. The institutions of government underwent great changes under the Frankish monarchs.

Kingship. Kingship was the basic institution in the Merovingian kingdom. Since Clovis' reign, the power of the king had extended not only over a tribe or tribes but also over a territory inhabited by Germans of divergent backgrounds and by Gallo-Romans as well. The king exercised power with legal limitations, which, when violated, led to efforts to reestablish political equilibrium by means of civil war, assassination, and an appeal to God and the saints. Royal power was dynastic and patrimonial. The Frankish kings successfully eliminated the Germanic practice of the magnates electing the king (the Frankish king was content to present himself to the magnates who acclaimed him) and accepted the hereditary principle as a personal right. The kings partitioned the kingdom at each succession. Royal power also had a sacred aspect; under the Merovingians the external sign of this was long hair.

Changes in the Frankish monarchy

The nature of the Frankish monarchy was profoundly changed during the Carolingian epoch. When Pepin III the Short restored the office of king in his own name, he had himself consecrated first by a bishop and then by the pope. This rite, of biblical origin, had already been adopted by the Visigoths; it gave Christian legitimacy to royal authority because it reinforced the religious character of the monarchy and signified the king's receipt of special grace from God. The king was permitted to reign and was given stature above the common level because of this grace. Acclamation by the magnates became a pledge of obeisance to a king whom God had invested with power.

To this new royal status Charlemagne added the title of emperor, which had not been held by a ruler in the West since 476. He adopted this title as a concession to the situation at the time. It was conferred in the course of a ceremony that, in spite of the presence of the pope, was of secular character. Among the clerical ranks that formed the entourage of the new emperor, the revival of the empire was regarded as a magistracy conferred by God in the interests of western Christianity and the church; imperial authority was considered a kind of priesthood, and its bearer was obligated to lead and protect the faithful. This idea reached fruition under Louis the Pious, who introduced the ceremony of anointing as part of the investiture; this ritual furnished the pope with the means for gaining a noteworthy role in the designation of the emperor. His role increased because, in spite of all the efforts of the clergy to maintain the integrity of the empire, the imperial title was returned to Italy. Later Carolingian emperors were designated on the basis of the interests and actions of the papacy.

The central government. By the time of Clovis, the ancient Germanic assembly of free men participated only in the conduct of local affairs and was consigned largely to a military role. Within each kingdom, the king's court, of Roman imperial origin but adapted and modified by the Frankish sovereigns, encompassed domestic services (treasury, provisioning, stables, clergy), a bureau of accounts, and a military force. The court was presided over by three men—the seneschal, the count of the palace, and, foremost, the mayor of the palace, who also presided over the king's estates. They traveled with the king, who, while having various privileged places of residence, did not live at a fixed capital. Only under Charlemagne did this pattern begin to change; while not abandoning the itinerant life, Charlemagne nonetheless wished to make Aachen the centre of his state. It was there that he constructed a vast palace, based upon a Roman model.

Local institutions. Except in the north, which was divided into districts called *pagi,* the Merovingians continued to use the city (the Roman *civitas*) as the principal administrative division. A count was installed in each *pagus* and city (*urbs*) and delegated financial, military, and judicial authority. Groups of counts were occasionally placed under the authority of a duke, whose responsibilities were primarily military.

The development of institutions in the Carolingian age. The Carolingians contented themselves with refining their administrative system in order to strengthen royal control and to solve the problems posed by a large empire. The kingdom's cohesion was augmented by an oath of fidelity, which Charlemagne exacted from every free man (789, 793, 802), and by the publication of legislation—the capitularies—which regulated the administration and exploitation of the kingdom. In the most exposed border areas, local governments—the marches—were established.

In order to offset the power of the counts, the episcopate was given a central role in the administration. Charlemagne extended the use of the *missi dominici*—*i.e.,* envoys who also served as liaisons between the central government and local agents and who were responsible for keeping the latter in line. In order to strengthen his control over the population, Charlemagne attempted to develop intermediary bodies; he tried to use both vassalage and immunity as means of government—in the first instance by creating royal vassals and giving them public offices and in the second by controlling protected institutions such as monasteries and the Jewish community.

Predominance of the estate

Economic life. Agriculture was the principal economic activity, and during the entire Frankish age the great estate, inherited from antiquity, remained the predominant characteristic of rural life. These estates were the residence and principal source of income of the aristocracy. The estates appear to have long been placed under cultivation by servile labour, which was abundant at the time. The heavy work was done with the assistance of day labourers. A portion of the land, however, was given to the tenants—the *coloni*—who were compelled to pay annual charges. With the decline in slavery at the end of the Merovingian era, the number of tenancies was increased, and tenants were compelled to render significant amounts of labour to cultivate land held directly by the lord. This evolved into a bipartite system, which was not adopted throughout the Frankish empire but was most widespread between the Loire and the Rhine.

Trade. Despite the Islāmic conquests, Mediterranean commerce did not decline abruptly. In Gaul, such goods as papyrus, oil, and spices were imported from the East, and there were numerous colonies of Syrians. Currency continued to be based on the gold standard, and imperial units were still used. All signs, moreover, point to the existence of manufacturing for trade (marble from Aquitaine, Rhenish glass, ceramics).

During the Carolingian age, Mediterranean trade no longer occupied a primary place in the economy. The adoption of a new monetary system based on silver along with a reduction in the number of Oriental goods and merchants are signs of the change. After the 7th century, trade among the countries bordering the English Channel and the North Sea and in the Meuse valley increased steadily. The Scandinavians, with their great commercial centres at Birka in Sweden and Hedeby in Denmark, were both pirates and traders; they established new contacts between East and West.

In addition to this large-scale commerce there was agriculturally based local trade. The number of markets increased, and market towns began to appear alongside the former Gallo-Roman cities, which survived as fortresses and population centres and served as the basis for religious organization and political administration.

Royal sources of income

Frankish fiscal law. The Frankish fiscal system reflected the evolution of the economy. Frankish kings were unable to continue the Roman system of direct taxation of land as the basis for their income. Their principal sources of income were the exploitation of the domains of the fisc (royal treasury), war (booty, tribute), the exercise of power (monetary and judicial rights), and the imposition of *telonea* (taxes collected on the circulation and sale of goods), the number of which was increased by the kings.

The church. The episcopate and the diocese were practically the only institutions to survive the collapse of Roman imperial power largely unchanged. During the German conquest many bishops played important roles in defending the population. During the Frankish era bishops and abbots occupied a socially prominent position because of both their great prestige among the people and their landed wealth.

Institutions. The organization of the secular church took its final form under the Merovingian and Carolingian kings. The administrative bodies and the hierarchy of the early Christian church were derived from institutions existing during the late Roman Empire. In principle, a bishop was responsible for the clergy and faithful in each district (*civitas*). The bishop whose seat was in the metropolitan city had preeminence and was archbishop over the other bishops in his archdiocese. The monarchy dominated the church. Kings most often appointed bishops from among their followers without regard for religious qualifications; the metropolitan see was often fragmented in the course of territorial partitions and tended to lose its importance, and the church in Francia tended to withdraw more and more from papal control despite papal attempts to reestablish ties.

The first Carolingians reestablished the ecclesiastical hierarchy. They restored the authority of the archbishops and established cathedral chapters so that the clergy living around a bishop were drawn into a communal life. They also maintained the right to nominate bishops, whom they considered agents of the monarchy.

During the 4th and 5th centuries success at converting the countryside made it necessary for the bishops to divide the dioceses into parish churches. Initially there was a limit of between 15 and 40 of these per diocese. In the Carolingian era they were replaced by small parish churches better suited to the conditions of rural life.

Monasticism. Monasticism originated in the East. It was introduced in the West during the 4th century and was developed in Gaul, mainly in the west (St. Martin of Tours) and southeast (St. Honoratus and St. John Cassian). The earliest monasteries lacked a communal form of discipline; eremitism was widespread. In the 6th century, monasteries throughout Gaul increased, and efforts were made to impose rules. The Irish monasticism introduced by St. Columban (*c.* 543–615) was influential in the 7th century, but it was later superseded by the Benedictine rule, introduced from Italy. The monasteries suffered from the upheavals affecting the church in the 8th century. The Carolingians attempted to reform the monasteries. Louis the Pious, acting on the advice of St. Benedict of Aniane, imposed the Benedictine rule; it became a characteristic feature of western monasticism. The Carolingians, however, continued the practice of having lay abbots.

Education. In the 6th century, especially in southern Gaul, the aristocracy and, consequently, the bishops drawn from it preserved an interest in traditional classical culture. Beginning in the 7th century, the Columbanian monasteries insisted on the study of the Bible and the celebration of the liturgy. In the Carolingian era these innovations shared the focus of education with works of classical antiquity.

Religious discipline and piety. One characteristic of the church in the 6th century was frequent councils to settle questions of doctrine and discipline. The conciliar institution declined, leading to liturgical anarchy and a moral and intellectual crisis among the clergy. Charlemagne and Louis the Pious attempted to impose a uniform liturgy, inspired by the one used at Rome. They also took measures to raise the standard of education of both clerics and the faithful.

The cults of saints and relics

The development of the cults of saints and relics was an important part of religion. Commerce in relics developed, with Rome as one of the key centres. The number of pilgrimages increased. The desire on the part of the faithful to be buried near relics changed funeral practices. Ancient cemeteries were abandoned, and burials in or near churches (burials *ad sanctos*) increased.

The influence of the church on society and legislation. The progressive Christianization of society influenced Frankish institutions significantly. The introduction of royal consecration and the creation of the empire afforded the clergy an opportunity to elaborate a new conception of power based on religious principles. The church was involved in trying to discourage slavery and in ameliorating the legal condition of those enslaved. It was during the Carolingian period that, in reaction to the polygamy practiced in German society, Christian doctrines of marriage were more strictly formulated.

Merovingian literature and arts. During the entire 6th century many writers, inspired by classical tradition, produced works patterned on antique models; such writers included Sidonius Apollinaris (d. *c.* 488), Gregory of Tours (d. 594/595), and Venantius Fortunatus (d. *c.* 600). The writing of saints' lives—hagiography—became the most widespread literary genre of the period. Nevertheless, the standard of literature continued to decline, becoming more and more conventional and artificial. The use of popular Latin became more common among writers.

Religious architecture remained faithful to the early Christian model (churches of basilican type, baptisteries, and vaulted mausoleums with central plans). Because of the development of the cult of saints and the practice of burying *ad sanctos,* mausoleums became common in churches. As had been the case in antiquity, marble was the principal sculptural material. In the Pyrenees, sculptors produced capitals and sarcophagi in antique style, which they exported throughout Gaul; these workshops reached their zenith in the 7th century. The development of the art of metalwork (fibulae, buckles) was another characteristic of the Merovingian age. Germanic craftsmen adapted Roman techniques (*e.g.,* cloisonné and damascene work). A new aesthetic standard, characterized by the play of colour and the use of stylized motifs, eventually predominated.

The Carolingian renaissance

Carolingian literature and arts. A renaissance movement occurred during the general renovation undertaken by the Carolingian monarchs; they supported the movement as an aid to religious reform and as a means to enhance their own prestige. The origins of the movement lie deep in the 7th and 8th centuries. It received great stimulus, however, when the growth of the empire brought the Franks into contact with lands where a higher cultural standard had been maintained and the antique tradition was still viable (Ireland, Italy, and Anglo-Saxon England). The Carolingian sovereigns attracted foreigners (*e.g.,* the Anglo-Saxon Alcuin and the Lombard Paul the Deacon) who played a decisive role in this renewal.

After raising the standard of the clergy, Charlemagne assembled a group of scholars at his court. Although, contrary to legend, there was no formal school established in the imperial palace, there were numerous schools opened in the vicinity of churches and monasteries. An attempt was also made to reform handwriting. Research was carried on simultaneously under the auspices of several monastic centres (most notably Tours) for the purpose of standardizing writing; this effort resulted in the adoption of a regular and uniform script (Carolingian minuscule). Improved teaching and a desire to imitate classical antiquity helped to revivify the Latin used by writers and scribes.

The imperial court was a hub of intellectual life. From it came works dedicated to the glory of the emperors, which were modeled on the examples of ancient authors. Important works include Einhard's *Vita Karoli Magni* (*Life of Charlemagne*), the Astronomer's *Vita Hludowici impera-*

toris (*Life of Louis the Pious*), Nithard's *Historiarum libri IV* (*History of the Sons of Louis the Pious*), and Hincmar's *De ordine palatii* ("On the Government of the Palace").

Regional cultures

Beginning in the mid-9th century, however, the kingdoms formed from the partitions of the empire saw a renaissance of regional cultures. The fact that the Oath of Strasbourg was drawn up in Romance and German is an early indication of this development. There is a striking contrast between the *Annales Bertiniani* ("Annals of St. Bertin"), written at the court of Charles the Bald, and the *Annales Fuldenses* ("Annals of Fulda"), written at the principal intellectual centre in Francia Orientalis. They are, respectively, the western and eastern narratives of the same events.

Some of the great imperial monuments erected during the Carolingian age (palace of Ingelheim, palace of Aachen) reveal the permanence of ancient tradition in their regular plans and conception. The churches were the subjects of numerous architectural experiments; while some were constructed on a central plan (Germigny-des-Prés, Aachen with its internal octagon shape), most remained faithful to the traditional T-shape basilican type. Liturgical considerations and the demands of the faith, however, made certain modifications necessary, such as crypts on the east or a westwork, or second apse on the west. These church buildings afforded architects an opportunity to make experiments in balancing the arches. The extension of the vaults over the entire church and the more rational integration of the annexes and church proper gave rise to Romanesque architecture.

The buildings of the period were richly decorated with paintings, frescoes, painted stucco, and mosaics in which figural representation increasingly replaced strictly ornamental decoration. North Italian ateliers were popularizing the use of interlace (*i.e.,* ornaments of intricately intertwined bands) in chancel decoration. Sumptuary arts became more common, especially illumination, ivory work, and metalwork for liturgical use (reliquaries).

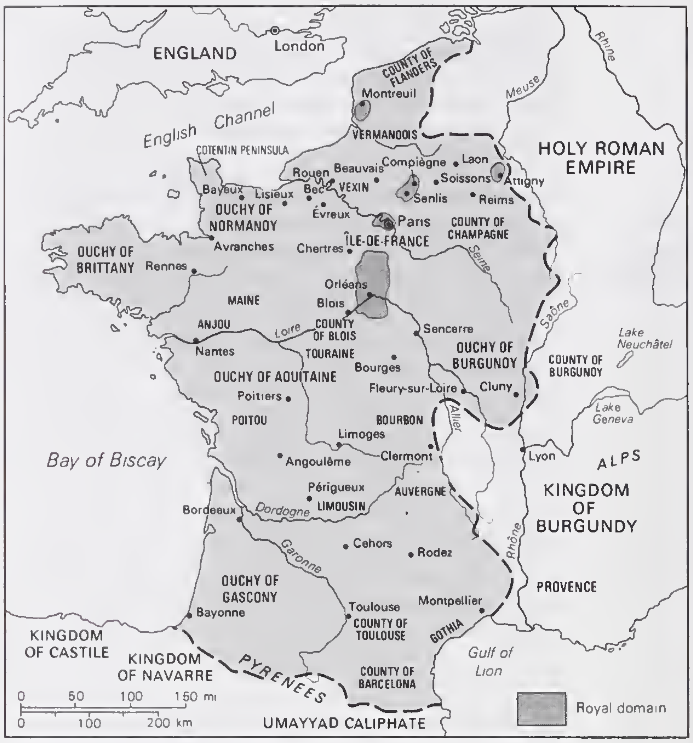
(G.Fo./B.S.Ba.)

The emergence of France

From the 9th to the 11th century the peoples and lands dominated by west Frankish kings were transformed. The Carolingian protectorate of local order collapsed under the pressures of external invasions and internal usurpations of power. Growing populations and quickening economies were reorganized in principalities whose leaders struggled to carry on the old programs of kings, bishops, and monks; one of these lands, centred on the Paris-Orléans axis and later known as the Île-de-France, was the nucleus of a new dynastic kingdom of France. This kingdom may be spoken of as Capetian France (the first king of the new dynasty having been Hugh Capet), but it was not until the 13th century that this France came to approximate the modern nation in territorial extent. The emergence of a greater France as a social and cultural entity preceded the political expansion of Capetian France; already in the 12th century crusaders, when speaking of "Franks" from Romance-speaking lands, meant something like "Frenchmen," while the persistence of old boundaries between populations of Romance and Germanic speech perpetuated the idea of a greater west Frankland.

FRENCH SOCIETY IN THE EARLY MIDDLE AGES

Social and political order. A foremost circumstance of the later 9th and 10th centuries was the inability of the west Frankish kings to keep order. The royal estates that had theretofore supported them, mostly in the north and east, were depleted through grants to retainers uncompensated by new acquisitions. Hindered by poor communications, the kings lost touch with lesser counts and bishops, while the greater counts and dukes strove to forge regional clienteles in fidelity to themselves. These princes (as they were called) were not rebels. More often allied with the king than not, they exercised regalian powers of justice, command, and constraint; it was typically they who undertook to defend local settlements and churches from the ravages of Magyars invading from the east, of Muslims on Mediterranean coasts, and of Vikings from northern waters.



France in 987.

Adapted from Ch. Petit-Dutaillis, *The Feudal Monarchy in France and England from the Tenth to the Thirteenth Century*, Barnes & Noble, Inc., and Routledge & Kegan Paul Ltd

Viking invasions

Of these invaders the Northmen, as contemporaries called the Vikings, were the most destructive. They raided landed estates and monasteries, seizing provisions and movable wealth. Striking as far inland as Paris by 845, they attacked Bordeaux, Toulouse, Orléans, and Angers between 863 and 875. From a base in the Somme estuary they pillaged Amiens, Cambrai, Reims, and Soissons. But they were drawn especially to the Seine valley. In 856–860 they laid waste the country around its lower reaches and repeatedly attacked Paris thereafter. Sometimes they were turned back by defenses but more often by payments of tribute. After 896 the invaders began to settle permanently in the lower Seine valley, whence they spread west to form the duchy of Normandy. Maritime raiding continued into the 10th century, then subsided.

Lords such as the counts of Flanders, Paris, Angers, and Provence were well situated to prosper in the crisis. They were often descended from or related to Carolingian kings. Adding protectorates over churches to their inherited offices, domains, and fiefs while acquiring other lordships and counties through marriage, they built up principalities that were as precarious as they were powerful. The lords tried to avoid dismemberment of the patrimony by limiting their children's right of succession and marriage, but it was only in the 12th century that these dynastic principles came to prevail in the French aristocracy. The princes, moreover, found it almost as hard as the kings to secure their power administratively. They exploited their lands through servants valued less for competence than for fidelity; these servants, however, were men who tended to think of themselves as lords rather than agents. This tendency was especially marked among the masters of castles (castellans), who by the year 1000 were claiming the power to command and punish as well as the right to retain the revenues generated from the exercise of such power. In this way was completed a devolution of power from the undivided empire of the 9th century to a checkerboard of lordships in the 11th—lordships in which the control of castles was the chief determinant of success.

Castellans

This fragmented polity was a feudal regime; at every level lords depended on the services of sworn retainers who were usually rewarded with the tenures of lordship called fiefs (*feudum-a*). In the 9th century fiefs were not yet numerous enough to undermine the public order protected by kings and their delegates. Indeed, fiefs were at first rewards for public service made from fiscal (royal) lands; this practice persisted in the south into the 11th century. By then, however, castles, knights, and knights' fiefs were multiplying beyond all control, resulting in a fracturing of power that few princes succeeded in reversing before 1100. Counts were unwilling to admit that their counties were fiefs or that they owed the same sort of allegiance to kings or dukes as their vassals did to them. Tainted with servility as well as with the brutality of needy knights on the make, vassalage was slow to gain respectability. The multiplication of fiefs was a violent process of subjugating free peasants and abusing churches.

Economic expansion. The population, still overwhelmingly agrarian, was growing from about 900 AD, probably most rapidly in the north, where more flexible schemes of crop rotation led to better harvests. Some peasants retained their independence, as in the Massif Central and the Pyrenees, but they were not much envied; still, small free properties nowhere entirely disappeared. Most peasants, however, were organized in subjection to lords—bishops, abbots, counts, barons, or knights—whose estates assumed diverse forms. In northern France lords typically reserved the proceeds of a domain worked by tenants, who had their own parcels of land to live on. Among such tenants were free men and slaves as well as a sizable intermediate estate of incompletely free villagers. Increasing productivity stimulated the improvement of roads and bridges, trade, and the growth of towns as well as competition for the profits of agrarian lordship. After about 1050, townspeople, especially merchants, sought to free themselves from the arbitrary lordship of counts and bishops, usually peaceably, as at Saint-Omer, but occasionally in violent uprisings, as at Le Mans and Laon.

The Cistercian order

Religious and cultural life. The Christian church was badly disrupted by the invasions. In Normandy five successive bishops of Coutances resided at Rouen, far from their war-torn district, which had lapsed into paganism. Elsewhere standards of clerical deportment declined, threatening the moral leadership by which Carolingian prelates had supported public order. Renewal came in two influential forms. First, monks in Burgundy and Lorraine were inspired to return to a strict observance of the Benedictine rule and thereby to win the adherence of lay people anxious to be saved. The monastery of Cluny, founded in 910 by a duke of Aquitaine with a bad conscience, not only stimulated a newly penitential piety that radiated beyond its walls but also encouraged reforms in other monastic houses. In the 11th century Cluny came to direct an order of affiliated monasteries that extended throughout France and beyond. Cluny's religious hegemony was challenged only in the 12th century with the rise of a yet more ascetic Benedictine observance, of which St. Bernard of Clairvaux (1090–1153) was the great proponent. Centred at Cîteaux in Burgundy (whence the appellation Cistercian), this movement combined ascetic severity with introspective spirituality and economic self-sufficiency. A newly personal devotionalism was diffused from monastic cloisters into lay society.

Second, the bishops, in the absence of royal leadership, renewed Carolingian sanctions against violence. The Peace of God was instituted in synods of southern France in the later 10th century. It was an effort, solemnized in ritual processions and oaths, to restrain the increasing number of knights from pillaging peasants and churches. It was supplemented from the 1020s by the Truce of God, which forbade fighting on certain days or during certain seasons of the year and which helped to mold a new conception of the knight as a Christian warrior prohibited from shedding the blood of Christian people. Warmly embraced by the Cluniac pope Urban II when he preached the First Crusade at Clermont in 1095, this idea contributed to a new ideal of knighthood as an honourable estate of Christian leadership. When young princes were dubbed to knighthood in the 12th century, they assumed a mode of respectability fashioned by the church; this eased the way for lesser knights to be recognized as nobles as well.

Peter Abelard

The growing wealth and stability of regional societies, as in Burgundy, Flanders, and Normandy, encouraged new impulses in the arts and letters. Cathedral churches supported scholars who revived the traditional curriculum of learning, stressing reading, writing, speaking, and computation. Fulbert of Chartres (d. 1028) was fondly remembered as a humane teacher by students who often became teachers themselves. A century later famous masters could be found at Laon and Paris as well as (probably) at Chartres, attracting young clerics to their lectures in swelling numbers. The Breton Peter Abelard (1079–1142) taught and wrote so brilliantly on logic, faith, and ethics that he established Paris' reputation for academic excellence. Traditional pursuits of contemplative theology and history gave way to new interests in logic and law. Men trained in canon and Roman law found their way increasingly into the service of kings, princes, and bishops.

Everywhere churches in Romanesque style were built, and they continued to be built in the south long after some architects, like Suger at Saint-Denis in the 1140s, introduced the new aesthetic of Gothic style. Lay culture found expression in vernacular epics, such as *The Song of Roland* in Old French, and the Provençal lyrics of southern France. These poems witness to diverse zones of linguistic evolution from spoken Latin; by the 12th century the *langue d'oïl* north of the Loire was broadly differentiated from the *langue d'oc* to the south. The cultural cleavage so marked ran deeper than language and was not entirely overcome by the spread of modern French from the *langue d'oïl.*

THE POLITICAL HISTORY OF FRANCE (C. 850–1180)

The fragmentation of political power meant that the kings of France were forced into rivalries, alliances, and conflicts with the princes, who were for many generations the real rulers of France.

Principalities north of the Loire. Outside of the dynastic royal domain (centred around Paris) the foremost northern powers were Flanders, Normandy, Anjou, Brittany, Blois-Champagne, and Burgundy.

The northernmost of these was Flanders, whose founder, Baldwin I Iron-Arm (862–879), managed not only to abduct the Carolingian king's daughter and marry her but also to win that king's approval as count of Ghent. His authority was consolidated under his son Baldwin II (879–918) and grandson Arnulf (918–965), the latter a violent and ambitious prince who undertook to restore the Flemish church as if he were an emperor. Fertile and precocious in trading activity, Flanders well supported such energetic lords; monks at Saint-Bertin and Ghent celebrated the dynastic feats of the counts.

In the time of Robert the Frisian (1071–93) efforts were made to systematize the count's lordship over castles as well as his fiscal rights, but the results fell short of giving the count effective sovereign power. When the foreign-born Charles the Good (1119–27) tried to pacify the county at the expense of lesser knightly families, he was murdered. Stability together with a new and centralized mode of fiscal accountancy was achieved by Thierry of Alsace (1128–68) and his son Philip (1163–91). Toward 1180 Flanders was a major power in northern France.

Normandy

The duchy of Normandy was created in 911, when the Viking chieftain Rollo (Hrolf) accepted lands around Rouen and Evreux from King Charles III the Simple. With its pastures, fisheries, and forests, this old land was a rich prize, and Rollo's successors extended their domination of it aggressively. Early Norman history, however, is more obscure than Flemish, lacking the records that only Christian clerics could write. The acquisitions of the second duke of Normandy, William I Longsword (927–942), were threatened when he was murdered by Arnulf I of Flanders in 942. It was only in the reign of his son Richard I (942–996) that something like administrative continuity based on succession to fiscal domains and control of the church was achieved. The dukes (as they then came to be styled) allied with the ascendant duke Hugh Capet had little to lose from the latter's accession to the kingship in 987; it was at this time that a new Norman aristocracy in ducal control took shape. Under Robert I the Devil (1027–35) agrarian and commercial prosperity favoured the multiplication of castellanies and knights, and Duke William II (1035–87; William the Conqueror) had to put down a dangerous rising of Norman barons and castellans in 1047 before proceeding, surely in deliberate consequence, to establish a firmly central control of castles that was without precedent in France. His conquest of England in 1066 made William the most powerful ruler in France. At the same time knights from lesser elite families in Normandy were establishing territorial lordships in southern Italy.

Norman ducal lordship was crude but effective. Under Henry I (1106–35) a unified exploitation of patronage, castles, and revenues was developed for the kingdom of England and the duchy of Normandy alike. Normandy passed to Henry's son-in-law Count Geoffrey of Anjou in 1135 and to his grandson Henry II (1154–89), in whose time it became the heartland of an Angevin dynastic empire.

Anjou, in the lower Loire valley, was among the lands delegated to Robert the Strong in 866. In the 10th century a series of vigorous counts established a dynastic patrimony that expanded under the great Fulk Nerra (987–1040) and his son Geoffrey Martel (1040–60) to include Maine and Touraine. Strategically situated, this principality prospered in its early times of external danger, but it was surrounded by aggressive dynasts; the control of castles and vassalic fidelities were the count's somewhat precarious means of power.

Brittany, to the west of Anjou and Normandy, was set apart by its strongly Celtic tradition. It achieved identity in the 9th century under the native leader Nomenoë, who seized Nantes and Rennes in defiance of Charles the Bald. His successors, badly battered by the Vikings, were recognized as dukes in the 10th century but were unable to consolidate their power over lesser counts and castellans. With little more than an unenvied independence, the duchy persisted in the 12th century when a series of succession crises enabled King Henry II of England to subject it to the Plantagenet domains. Only after 1166 were the Bretons to feel the impact of systematic territorial administration.

The area around Blois, to the east of Touraine, had also been entrusted to Robert the Strong and remained in his family's hands until about 940, when Theobald the Old seized control of it and founded a line of counts of Blois. His successors, notably the fearsome Eudes II (996–1037), annexed the counties of Sancerre (1015) and Champagne (1019–23), thereby creating a principality comparable in strength to Flanders and more threatening to the king, whose patrimonial domains it encircled. A dynastic aggregate lacking natural cohesion, Blois-Champagne achieved its greatest strength under Theobald IV the Great (II of Champagne, 1125–52), who was a formidable rival of Kings Louis VI and Louis VII. The main lands were divided under his sons Theobald V (1152–91) and Henry (1152–81), themselves prestigious lords; and the Champagne of Henry the Liberal was among the richest, best organized, and most cultured French lands of its day.

Burgundy

Finally, there was Burgundy, to the south of Champagne (not to be confused with the old kingdom and the later imperial county of Burgundy), which first achieved princely identity under Richard the Justiciar (880–921). Defeating Magyars and Vikings as well as exploiting the rivalries of his neighbours, Richard was regarded (like his near contemporary Arnulf of Flanders) as virtually a king. Ducal power was contested and diminished thereafter, but it survived as the patrimony of a Capetian cadet family until 1361.

Thus, by the later 12th century, France north of the Loire consisted of several large principalities (some of them associated with the English crown), coexisting with each other and with the king, who struggled to impose his lordship on them.

The principalities of the south. South of the Loire emerged another set of lands: Provence, Auvergne, Toulouse, Barcelona, and Aquitaine.

Provence lay in what is now the southeastern corner of France; it was not part of the west Frankish domains. Included in the Middle Kingdom from 843, it passed to the kings of Burgundy after 879 and to the emperors in the 11th century. But it was local counts once again who won prestige as defenders against pillagers, in this case the Muslims, and profited from urban growth to establish a dynastic authority of their own. This authority was fractured in the early 12th century, when the houses of Barcelona and Toulouse secured portions by marriage; a cadet dynasty of Barcelona continued to rule the county until 1245.

The county of Barcelona, formed from a delegation of Frankish royal power in 878, came to dominate all other east Pyrenean counties in the 11th century. Prospering at the expense of the Muslims, Count Ramon Berenguer I (1035–76) reduced his castellans to submission (like his contemporary William in Normandy). His great-grandson Ramon Berenguer IV (1131–62) organized the strongest principality in the south. He and his successors acted as fully independent sovereigns, although the king of France retained a theoretical lordship over Barcelona until 1258.

Auvergne is the best example of a region whose masters failed to subordinate rival counts and castellans. Only a tradition of superior comital unity survived in the claims of two related counts whose patrimonies were absorbed by the crown in the 13th century.

Toulouse

Toulouse had been a centre of delegated Frankish power from the 8th century, but its pretension to princely status dated from 924, when Raymond III Pons (924–after 944) added control of coastal Gothia to that of Toulouse and its hinterland. Dynastic continuity, here as elsewhere, however, was badly interrupted, and none of the succeeding counts was able to organize a coherent lordship. Raymond IV of Saint-Gilles (1093–1105) acquired the crusader land of Tripoli (Syria), but he and his successors were weakened at home by conflicts with Barcelona and Aquitaine.

The duchy of Aquitaine might at first have seemed the most promising of all these principalities. A kingdom in the 9th century, it was reconstituted under William the Pious (d. 926) and again, more imposingly, under William V (995–1029), who was acclaimed as one of the greatest rulers of his day. Yet his power depended on lordships and alliances rather than on administration, and so the situation remained in the 12th century, when the vast but flabby duchy was conveyed by the marriages of its heiress Eleanor successively to the kings of France and England.

Of these principalities only Barcelona had achieved territorial cohesion and cultural unity by the later 12th century; it was then becoming known as Catalonia. The others, less toughened by external invasion and less resistant to the Cathari (or Albigensian) religious heresy from within, were vulnerable to an expanding Capetian monarchy.

The monarchy. The kingdom of France was descended directly from the west Frankish realm ceded to Charles the Bald in 843. Not until 987 was the Carolingian dynastic line set aside, but there had been portentous interruptions. The reunited empire of Charles the Fat (884–888) proved unworkable: the Viking onslaught was then at its worst, and the king proved incapable of managing defenses, which fell naturally to the regional magnates. Among these was Eudes, son of that Robert the Strong to whom counties in the lower Loire valley had been delegated in 866. Eudes's resourceful defense of Paris against the Vikings in 885 contrasted starkly with Charles the Fat's failures, and in 887 the west Frankish magnates deposed Charles and later elected Eudes king. In so doing they bypassed an underage grandson of Charles the Bald, also named Charles, who was crowned at Reims in 893 with the support of the archbishop there. Although gaining undisputed title to the crown upon Eudes's death in 898, Charles the Simple (as he was called) was unable to recover the undivided loyalty of his great men. Having ceded Normandy to a Viking band in 911 and having sought to reward the service of lesser men, he lost the crown in 922 to Eudes's brother Robert, who was killed in battle against Charles in 923. Thereupon Robert's son-in-law Raoul of Burgundy was elected king and Charles the Simple was imprisoned, to die in captivity in 929. Yet, when Raoul died in 936, the Robertian candidate for the crown, Robert's son Hugh the Great, stood aside for another Carolingian restoration in the person of Louis IV, son of Charles the Simple and called d'Outremer ("from Overseas") because he had been nurtured in England since his father's deposition. Louis IV acted energetically to revive the prestige of his dynasty, leaving the crown undisputed at his death in 954 to his son Lothaire (954–986). But Lothaire's dynastic resources were too seriously impaired to command the full allegiance of the magnates. When his son Louis V (986–987) died young, the magnates reasserted themselves to elect Hugh Capet king. This time, despite the survival of a Carolingian claimant, Charles of Lorraine, the dynastic breach was permanent.

Crisis of power

The election of 987 coincided with a more general crisis of power. The pillaging of Vikings gave way to that of castellans and knights; the inability of kings (of whatever family) to secure professions of fidelity and service from the mass of people in lands extending beyond a few counties shows how notions of personal loyalty and lordship were replacing that of public order. Just as castellans were freeing themselves from subordination to counts, so the monks claimed exemption from the supervision of bishops: in a famous case the bishop of Orléans was opposed by the learned Abbo of Fleury (d. 1004). There was a new insistence on the virtue of fidelity—and on the sin of betrayal.

Hugh Capet (987–996) and his son Robert (996–1031) struggled vainly to maintain the Carolingian solidarity of associated counts, bishops, and abbots; after about 1025 Robert and his successors were hardly more than crowned lords, and their protectorate was valued by few but the lesser barons and churches of the Île-de-France. Neither Henry I (1031–60) nor Philip I (1060–1108) could match the success (such as it was) of their rivals in Normandy and Flanders in subordinating castles and vassals to their purposes.

Yet even these relatively weak kings clung to their pretensions. They claimed rights in bishops' churches and monasteries far outside their immediate domain, which was concentrated around Paris, Orléans, Compiègne, Soissons, and Beauvais. Henry I married a Russian princess, whose son was given the exotic name of Philip; and the choice of Louis, a Carolingian name, for Philip's son was even more obviously programmatic. Louis IV (1108–37) spent his reign reducing the robber barons of the Île-de-France to submission, thereby restoring respect for the king's justice; he worked cautiously to promote the royal suzerainty over princely domains. It was a sign of newly achieved prestige that he secured the heiress Eleanor of Aquitaine as a bride for his son Louis VII (1137–80). But Louis VI was less successful in border wars with Henry I of Normandy; these conflicts became more dangerous when, upon the failure of her first marriage, Eleanor married Henry of Anjou, who came thereby to control lands in western France of much greater extent than the Capetian domains. Louis VII proved nonetheless a steady defender of his realm. He never relinquished his claim to lordship over the Angevin lands, and he allowed lesser men of his entourage the freedom to develop a more efficient control of his patrimonial estate. Not least, he fathered—belatedly, by Queen Adele of Champagne, his third wife, amid transports of relieved joy—the son who was to carry on the dynasty's work.

The early Capetian kings thus achieved the power of a great principality, such as Normandy or Barcelona, while harbouring the potential to reestablish a fully royal authority over the greater realm once ruled by Charles the Bald. The princes were their allies or their rivals; they sometimes did homage and swore fealty to the king, but they were reluctant to admit that their hard-won patrimonies were fiefs held of the crown. Royal lordship over peasants, townspeople, and church lands was for many generations a more important component of the king's power in France. It was exercised personally, not bureaucratically. The king's entourage, like those of the princes, replicated the old Frankish structure of domestic service. The seneschal saw to general management and provisioning, a function (like that of the mayors of the palace) with the potential to expand. The butler, constable, and chamberlain were also laymen, the chancellor normally a cleric. The lay officers were not agents in the modern sense; their functions (and incomes) were endowed rewards or fiefs, for which they seldom accounted and which they tended to claim as by hereditary right. In a notorious case, Stephen of Garland tried to claim the seneschalsy as his property and for a time even held three offices at once; but this abuse was soon remedied and taught caution to Louis VI and his successors. The chancellor drafted the king's decrees and privileges with increasing care and regularity. He or the chamberlain kept lists of fiscal tenants and their obligations on the lord-king's estates and in towns for use in verifying the service of provosts who collected the rents and profits of justice. But this service was hardly less exploitative than that of the household officers; the royal domain lagged behind the princely ones of Flanders and Normandy in the imposition of accountability on its servants. The abbot Suger of Saint-Denis (d. 1151), once a provost on his monastery's domains, was instrumental in furthering administrative conceptions of power in the court of Louis VII.

France, 1180 to *c.* 1490

FRANCE FROM 1180 TO 1328

The kings and the royal government. The French monarchy was greatly strengthened by Louis VII's successor, Philip II Augustus (ruled 1180–1223), who could claim descent from Charlemagne through his mother. Philip proved to be the ablest Capetian yet to reign. He was practical and clear-sighted in his political objectives; the extension of territorial power and the improvement of mechanisms with which to govern an expanded realm were his consistent policies. Perhaps it was not accidental that royal documents began to refer to the king of France (*rex Franciae*) instead of using the customary formula king of the Franks (*rex Francorum*) within a year or two of Philip's accession.

Territorial expansion

Philip Augustus. Philip's outstanding achievement was to wrest control from the Plantagenets of most of the domains they held in France. Intervening in struggles between Henry II of England and his sons, Philip won preliminary concessions in 1187 and 1189. He acquired strategic lands on the Norman borders following wars with Henry's sons, King Richard and King John (1196 and 1200). And when, in 1202, John failed to answer a summons to the vassalic court of his lord, Philip Augustus confiscated his fiefs. Normandy, invaded in 1204, submitted to the Capetian in 1208. Maine, Anjou, and Touraine fell rapidly (1204–06), leaving only Aquitaine and a few peripheral domains in the contested possession of England. By the Truce of Chinon (Sept. 18, 1214), John recognized the conquests of Philip Augustus and renounced the suzerainty of Brittany, although the complete submission of Poitou and Saintonge was to take another generation.

Philip's other acquisitions of territory, if less spectacular, were no less important for consolidating the realm. In the north he pressed the royal authority to the border of Flanders. Artois, which came under his control as a dowry with his first wife, was fully secured in 1212. Vermandois and Valois (1213) and the counties of Beaumont-sur-Oise and Clermont-en-Beauvais were annexed during his last years. On the southern limits of the Île-de-France Philip rounded out prior possessions in Gâtinais and Berry. Much of Auvergne, whose suzerainty had been ceded by Henry II in 1189, passed to royal control in 1214, while in the more distant south Philip extended his influence by gaining lordship over Tournon, Cahors, Gourdon, and Montlaur in Vivarais. As the reign ended, only Brittany, Flanders, Champagne, Burgundy, and Toulouse, among principalities later annexed, lay outside the royal domain.

Because the territorial expansion was accomplished through traditional means—dynastic, feudal, and military—the curial administration was, outwardly, little changed. Household officers such as the butler and the constable continued to function as in the past. But Philip Augustus was even more suspicious of the seneschalship and chancellorship than his father had been; he allowed both offices to fall vacant early in his reign, entrusting their operations to lesser nobles or to clerics of the entourage. Although their activity is obscure, some of these men were beginning to specialize in justice or finance. The curia as such, however, remained undifferentiated; characteristically, the committee of regents, appointed in 1190 to hold three courts yearly while the king was absent on crusade, was expected to function in both justice and administrative review on those occasions. Prelates and nobles of the curia also served as counselors; enlarged councils convened, at the king's summons, on festivals or when major political or military issues were contemplated.

Royal administration

Philip Augustus acted vigorously to improve the efficiency of his lordship. He was, indeed, practically the founder of royal administration in France. His chancery began to keep better records of royal activities. Documents were copied into registers before being sent out, and lists of churches, vassals, and towns were drawn up to inform the king of his military and fiscal rights. These lists replaced others lost on the battlefield of Fréteval·(1194), a disaster that may have hastened the adoption of a new form of fiscal accountancy. One may draw this conclusion because it is unlikely that the Capetians had previously troubled to record the balances of revenues and expenses in the form first revealed by a record of the year 1202. Its central audit was connected with other efforts to improve control of the domains dominated directly by the king. From early in his reign Philip appointed members of his court to hold periodic local sessions, to collect extraordinary revenues, to lead military contingents, and to supervise the provosts. The new officers, called bailiffs (*baillis*), at first had no determined districts in which to serve (they resembled the circuit commissioners of Angevin government, whose office may have been the model for the Capetian institution). From the outset the bailiffs were paid salaries; they were more reliable than the provosts, who, by the later 12th century, generally farmed the revenues. In the newly acquired lands of the west and south Philip and his successors instituted seneschals—functionaries similar to the bailiffs, but with recognized territorial jurisdiction from the start.

Philip Augustus' policy toward his conquered domains was shrewd. He retained the deep-rooted customs and administrative institutions of such flourishing provinces as Anjou and Normandy; indeed, the superior fiscal procedures of Normandy soon exercised perceptible influence on Capetian accounting elsewhere. On the other hand, to secure the loyal operation of provincial institutions, Philip appointed men of his own court, typically natives of the Île-de-France. It was a compromise that was to work well for generations to come.

Philip's relations with his subjects

The character of Philip's rule may likewise be deduced from his relations with the main classes of the population. A devoted son of the church, if not unswervingly faithful, he favoured the higher clergy in many of their interests. He opposed the infidels, heretics, and blasphemers; he supported the bishops of Laon, Beauvais, Sens, and Le Puy (among others) in their disputes with townspeople; and he granted and confirmed charters to monasteries and churches. Yet he was more insistent on his rights over the clergy than his predecessors had been. He required professions of fidelity and military service from bishops and abbots, cited prelates to his court, and sought to limit the jurisdiction of ecclesiastical courts. He supported papal policies or submitted to papal directives only to the extent that these were consistent with his temporal interests. His reserved support of crusades and his notorious rejection of Queen Ingeborg were cases in point (see below).

Toward the lay aristocracy, Philip Augustus acted energetically as suzerain and protector. Indeed, no Capetian was more fully the "feudal monarch." His war with John resulted from a breach of feudal law and was fought with feudal levies. He regarded Flanders and Toulouse as well as Normandy as fiefs held of the crown. As with ecclesiastical vassals, Philip insisted upon the service due from fiefs; he exploited the feudal incidents, notably relief and wardship; and he required his vassals to reserve their fealty for himself alone. He extended his influence by entering into treaties (*pariages*) with minor lords, often distant ones; and, by confirming the acts of nobles in unprecedented numbers, he recovered the force of the royal guarantee.

The policy toward the lesser rural and urban populations was to increase their loyalty and contribution to the crown without significantly reducing their dependence on the king and other lords. Philip offered his protection to exploited villages, and, especially during his early years, he confirmed existing "new towns," extended their privileges to other villages, and otherwise favoured peasant communities. Townsmen, notably those in semiautonomous communes, gained confirmation of their charters; and the king created some new communes. Most of the latter were located in strategic proximity to the northern frontiers of the expanded royal domain; this fact, together with the obligations of service and payment specified in the charters, suggests that military motives were paramount in these foundations. More generally evident in these charters, as in others, was the desire to gain the political fidelity of a prospering class. At Paris Philip Augustus acted as did no other local lord to promote the civic interest: improving sanitation, paving streets, and building a new wall. Parisian burghers financed and administered these projects; they were associated in the fiscal supervision of the realm when the king went on crusade, but they were not favoured with a communal charter.

Appanages

Louis VIII. The reign of Louis VIII (ruled 1223–26) had an importance out of proportion to its brevity. It was he (this frail husband of the formidable Blanche of Castile and father of famous sons) who first brought Languedoc under the crown of France and who inaugurated the appanages—grants of patrimonial land to members of the royal family or royal favourites—thereby creating a familial condominium through which the expanded France of later generations was to be governed. The conquest of Languedoc, following the Albigensian Crusade (against heretics in southern France) that was only tepidly supported by Philip Augustus, was not complete until the 1240s, but the royal seneschalsies of Beaucaire and Carcassonne were already functioning when Louis VIII died. And it was in keeping with that ruler's will of 1225 that the great appanages passed to his younger sons as they came of age—Artois to Robert in 1237; Poitou, Saintonge, and Auvergne to Alphonse in 1241; and Anjou and Maine to Charles in 1246.

Louis IX. The real successor to Philip Augustus, however, was his grandson, Louis IX (ruled 1226–70), in whose reign were fulfilled some of the grand tendencies of prior Capetian history.

Louis IX, who was canonized in 1297, is the best-known Capetian ruler. He impressed all who came in touch with him, and the records of his reign—anecdotal and historical as well as official—leave no doubt that he commanded affection and respect in a combination and to an extent that were unique. He regarded himself as a Christian ruler, duty-bound to lead his people to salvation. He led by example, precept, and correction. He earned a reputation for fairness and wisdom that enabled him to rule as absolutely as he wished; only with the crusade, perhaps, did his judgment falter. His reign was marked by consolidation, maturation, and reform rather than by innovation.

Louis IX's territorial acquisitions

In his early years baronial revolts, supported by Henry III of England, were put down by the regency, headed by the queen mother, with singular firmness and skill. Poitou and Saintonge remained restive largely because of the stubborn machinations of Isabella of Angoulême (King John's widow); it was only in 1243, after a revolt planned to coincide with an uprising in Languedoc, that the feudal adjudication of 1202 was fulfilled in Aquitaine. The revolt of Raymond Trencavel, dispossessed heir to the viscounty of Béziers, halfheartedly supported by Raymond VII of Toulouse, was no more successful; its failure resulted in the vindictive destruction of the petty nobility of Languedoc, and many fiefs thereupon passed to the crown. In 1239 a childless count of Mâcon sold his domains to the king.

Such were the principal territorial acquisitions of Louis IX; the balance of his work, however, was to be affected further by three characteristic events. First, despite his victory of 1243, Louis remained disposed to compromise with Henry III; in the Treaty of Paris (December 1259) Henry regained feudal title to lands and reversionary rights in Guyenne in exchange for renouncing all claims to Normandy, Anjou, Maine, Touraine, and Poitou. Similarly, by the Treaty of Corbeil (May 1258) Louis himself had abandoned ancient claims to Catalonia and Roussillon in exchange for the renunciation of Barcelona's rights in Gévaudan and Rouergue. Meanwhile, upon the death of Raymond VII in 1249, the county of Toulouse had passed to Raymond's son-in-law, Alphonse of Poitiers, who proceeded to govern it as effectively as his appanage lands; and when he and his wife died without issue in 1271, their enormous inheritance reverted to the royal domain.

The ancient household administration died out in the 13th century. Offices such as the chancery and treasury became more specialized and bureaucratic, while the greater advisory personnel formed a fluctuating corps of reliable favourites: bishops, abbots, and minor nobles of the old Capetian homelands. The counselors, meeting in diverse political and ceremonial capacities, continued to assemble with other prelates and barons during festivals or ad hoc. But the fiscal and judicial activities of the court were growing in volume and technicality. Ordinary revenues expanded apace with the royal domains; taxes ceased to be exceptional. Toward 1250, judgments of the curia began to be recorded centrally; and the judicial sessions, now often called *parlements*, derived an ever-expanding jurisdiction from the king's repute.

Meanwhile, a real local administration evolved as the bailiffs and seneschals became well established in territorial circumscriptions. Complaints arose when these men, and more particularly their subordinate officers, abused their powers for personal profit or the king's. Commissions of investigation, first appointed in 1247, provided means for redress; and these investigators continued to function after Louis returned from his first crusade in 1254.

Domestic policies of Louis IX

Although previous rulers had legislated on occasion, Louis IX was the first to express his will regularly in statutory form. A great ordinance for administrative reform in 1254 resulted from the remedial inquiries. In other enactments, characteristically moral and authoritarian, Louis sought to curb private warfare (about 1258) and to promote the use of royal money while limiting that of baronial (1263–65).

Toward the clergy Louis IX manifested a sympathy born of conservatism and exceptional piety, but he was nonetheless a firm master. He opposed efforts to expand clerical jurisdictions. During his later years he supported papal taxes on the clergy for the crusade, although in the 1240s he had joined his clergy in opposing papal preferments and impositions for a war against the emperor Frederick II. The lay nobles found Louis IX a frustrating ruler. Sharing few of their values, he consistently tried to limit their ability to cause disorder. He allowed royal officials to encroach on baronial jurisdiction in many cases, and he welcomed appeals from baronial judgments. On the other hand, he respected such rights as were sanctioned by provincial custom and was less forceful in exploiting feudal relationships than his grandfather had been.

The royal interest in order and justice was especially beneficial to townspeople and peasants, who had most suffered from exploitative agents and private war. Louis IX confirmed municipal charters, but he also taxed the towns heavily. When oligarchical urban governors mismanaged finance to the disadvantage of the lower classes as well as the king, he moved energetically (1259–62) to place the fiscal administration of 35 communes directly under the crown. A crusade of peasants known as the Pastoureaux (1251) was inspired by loyalty to the king, then in trouble in the Holy Land; when its impulse was dissipated in agitation against the propertied classes, the regent had it suppressed.

Later Capetians. Louis IX was succeeded by his son, Philip III (ruled 1270–85); his grandson, Philip IV the Fair (ruled 1285–1314); and three great-grandsons, Louis X (ruled 1314–16), Philip V (ruled 1316–22), and Charles IV (ruled 1322–28). The greatest of these last Capetian reigns was that of Philip IV. Worldly and ambitious, yet pious and intelligent, he was less accommodating than his forebears and more devoted to his power than to his reputation. He brought the monarchy to a degree of coordinated strength it was not again to have in the Middle Ages. But, in so doing, he strained the resources and patience of his subjects. His sons had to give in to the demands of a tired country but did so without abandoning their father's objectives. When Charles IV died without a male heir in 1328, as his brothers had done before him, the royal succession was claimed by a collateral Capetian family.

Further territorial consolidation

The reigns of the later Capetian kings were marked by further territorial consolidation. Marrying his son to the heiress of Champagne and Navarre in 1284, Philip III prepared the way for a reversion no less important than that of Toulouse (1271). Philip the Fair secured the heiress to the county of Burgundy for his son Philip in 1295 and annexed southern Flanders and Lyon in 1312. Smaller acquisitions, cumulatively of great importance, resulted from purchase: the counties of Guînes (1281), Chartres (1286), La Marche and Saintonge (1308); the viscounties of Lomagne and Auvillars (1302) and La Soule (1306); and a number of untitled lordships.

Through treaties, Philip the Fair extended his jurisdiction into the ecclesiastical principalities of Viviers, Cahors, Mende, and Le Puy. With his greatly expanded domain, the king could assert unprecedented authority everywhere in France. Yet it does not appear that territorial policy as such had changed. Appanages were still to be granted and to be recovered by the later Capetians. The monarchs continued to do without Brittany, Burgundy, and many lesser lordships, which did not prevent them from legislating for these lands along with the rest.

Government became more engrossing, specialized, and efficient. Although the royal curia continued to exist as an aggregate of favourites, magnates, prelates, and advisers, its ministerial element—comprising salaried officers serving at the king's pleasure—functioned increasingly in departments. The small council acquired definition from an oath first mentioned in 1269. Parlement, its sessions lengthening under a growing burden of cases, was divided into chambers of pleas, requests, and investigations (1278), and its composition and jurisdiction were regulated. Older provincial tribunals, such as the Norman Exchequer and the Jours of Troyes, became commissions of Parlement. While the direction of finance was left with the council, the Chamber of Accounts, apart from the treasury, was organized to audit accounts. Council and chamber as well as Parlement developed appropriate jurisdiction, and all three bodies kept archives. The chancery, serving all departments, remained in the hands of lesser functionaries until 1315, when Louis X revived the title of honour.

Local administration was marked by the proliferation of officers subordinate to the bailiffs and seneschals. The chief judge (*juge-mage*) assumed the seneschal's judicial functions in the south; receivers of revenues, first appearing in Languedoc, were instituted in the bailiwicks at the end of the 13th century. Commissions of investigation continued to traverse the provinces under the later Capetians, but all too often they now functioned as fiscal agents rather than as reformers.

Royal authority of Philip the Fair

Many of the officers who served Philip the Fair were laymen, and many were lawyers. Impressed with the power they wielded, they promoted loyalty to the crown and a conception of the royal authority approaching that of sovereignty. Without claiming absolute power for the king, they thought in terms of his "superiority" over all men within national boundaries now (for the first time) strictly determined; and they did not hesitate to argue from Roman law that, when the "state of the kingdom" was endangered, the monarch had an overriding right to the aid of all his subjects in its defense. While this doctrine, in a notorious case, was made a justification for imposing on the clergy, the later Capetians did not lose the religious mystique they had inherited from their predecessors' efforts in Christian causes. Even as political loyalties were being engrossed by the lay state, the "religion of monarchy" derived impetus from the fervent utterance of those who saw in Philip the Fair a type of Christ or the ruler of a chosen and favoured people.

It was in the requirements of war and finance that the claims of the monarchy found most concrete expression. In the 1270s, for his campaigns in the south, Philip III requested military aid from men theretofore exempt from such service. Philip the Fair, renewing these demands for his wars in Gascony and Flanders, went so far as to claim the military obligation of all free men as the basis for taxing personal property. The most persistent and lucrative taxation after 1285 was that imposed on the clergy, generally in the form of tenths and annates; sales taxes, customs, tallages on Jews and foreign businessmen, and forced loans likewise supplemented older revenues of the domain to support increased administrative expenses as well as costs of war. The most unpopular fiscal expedients were the revaluations of coinage after 1295, by which the king several times increased the profits of his mints to the confusion of merchants and bankers. The imbalance between ordinary resources and the needs of an expanding government became chronic at the end of the 13th century. Yet, in spite of the statist arguments of their lawyers, none of the later Capetians was moved to regard taxation as an established and justified requirement of a national government.

Philip's use of councils and assemblies

Such restraint is one reason why, with momentary lapses, the strongest of the later Capetians was not regarded as an arbitrary ruler. Philip the Fair revered St. Louis (Louis IX) as much as did his people; like Louis he took counsel from a relatively few unrepresentative persons. But, when Philip's own policies broke with the past, he resorted to great councils and assemblies, not so much to commit the nation as to justify his course. Whether a tax was sanctioned by custom or not, even if approved by assembled magnates or townsmen, he had it negotiated—reexplained and collected—in the provinces and localities. Large central assemblies in 1302, 1303, 1308, and 1312 met to enable the king and his ministers to arouse political support for his measures against the pope or the Knights Templars (see below).

Among these assemblies were the earliest national assemblies that included representatives of towns and villages, a fact that has caused historians to classify them with the Estates-General. Under Philip the Fair and his sons, however, these convocations were not yet understood to be representative of the estates of society; only when Philip V began to summon northern and southern men separately to deliberate on fiscal matters were the estates (which made up the Estates-General) in any way anticipated. Almost simultaneously the provincial Estates were foreshadowed in the petitions of magnates and towns in several regions for relief from administrative violations of traditional privilege; but the resulting charters of 1314–15 were poorly coordinated and reactionary, and they did little to limit royal power, although the fiscal rights later claimed by the Estates of Normandy could be traced to the Norman Charter of 1315.

If the policies of Philip the Fair evoked the complaint of all classes of people, it was because he had favoured none in particular; in fact, except in war and finance, the later Capetians may be said to have maintained a traditional politics toward both the nobles and towns. With the church, however, it was otherwise. Philip the Fair's insistence on taxing the clergy for defense led immediately to his conflict with Pope Boniface VIII. The latter, in the bull *Clericis laicos* (1296), issued on the protest of both French and English prelates, forbade the payment of taxes by clergymen to lay rulers without papal consent; but Philip's anger and his political arguments divided the clergy, and the pope soon found it necessary to abandon his position.

The quarrel was renewed in 1301, when the king and the magnates accused the bishop of Pamiers of treason and heresy. Boniface not only revoked the concessions of 1297 but rebuked Philip for seizing clerical property and debasing the coinage, among other things, and he summoned French prelates to Rome to proceed with a reform of the kingdom. Once again the clergy was split; many bishops and abbots attended an assembly at Paris in 1302 where they joined men of the other estates in addressing a remonstrance to the pope. A year later the king adopted rougher tactics: in June 1303 many prelates acquiesced in a scheme to try the pope before a general council, and in September the king's envoy Guillaume de Nogaret and his accomplices seized Boniface at Anagni. When the aged pope died a month later, his audacious policy collapsed entirely. The Gascon pope Clement V (reigned 1305–14) moved the Holy See to Avignon, and a mass of his compatriots were appointed cardinals.

Destruction of the Templars

With this pliant pontiff, the way was cleared for the strangest act of violence of the reign of Philip the Fair—the destruction of the Knights Templars. Founded in the 12th century in support of the Crusades, the French Templars by about 1300 were aged men whose privileges seemed poorly justified after the fall of the Holy Land. But Philip's case was pressed relentlessly beyond the evidence brought forth; the king had to resort to propaganda in a vast assembly held at Tours in 1308, and it was only in 1312 that the pope, his scruples overcome by expediency, dissolved the order. The Templars' possessions were given to the Hospitalers, and their last dignitaries were imprisoned or executed.

Foreign relations. France assumed a more active role in the politics of Christian Europe from the end of the 12th century. Philip Augustus led French contingents on the most fully international of the great Crusades (1190–91), although, having once demonstrated his energy in that work of piety, he could not afterward be persuaded to renew his vow. He preferred, through dynastic schemes and opportunism, to exploit his rivalry with the Plantagenets. His ambition seems to have embraced England as early as 1193, when he married Ingeborg, whose brother, the king of Denmark, had an old claim to the throne of England. When Philip, for private reasons, repudiated Ingeborg the day after the wedding and sought to have the marriage annulled, she and her brother appealed to the pope; her case, punctuated by reconciliations with Philip dictated more by policy than by sentiment, dragged on through the pontificate of Innocent III.

Meanwhile, in 1200, Philip's son Louis married Blanche of Castile, granddaughter of Henry II, through whom another claim to England was heralded. Louis's career as prince was marked by aggressive designs against King John. Innocent III was prepared to recognize Louis as king of England in 1213; and the policy was dropped only after Louis's abortive invasion of 1216–17.

It was in the play of rival coalitions that Philip Augustus had his greatest diplomatic anxiety and success. Philip countered John's alliance with Otto of Brunswick, his nephew and claimant to the empire, by supporting a second claimant, Philip of Swabia. When Otto became emperor in 1209 and the counts of Flanders and Boulogne were alienated from their Capetian suzerain, Philip found himself seriously threatened in his northern heartlands. John's desire to avenge the loss of his French fiefs finally prompted him to act in 1214; he himself led a force from the west, and his major allies marched on Paris from the north. Philip Augustus met the allied forces at Bouvines in July 1214 and won a decisive victory. As John retreated and his coalition collapsed, there could be no doubt that Capetian France had achieved hegemony in Christian Europe.

Diplomacy of Louis IX

Louis IX acted astutely, though in ways unlike his grandfather's, to preserve the prestige of France. His treaties with Aragon and England, designed to extend and secure his domains, resulted from a cordiality better appreciated abroad than by the royal counselors. From Navarre and Lorraine as well as from within the realm were brought disputes for his judgment; and in the Mise of Amiens (1264) Louis responded to the appeal of Henry III and the English barons to pronounce on the validity of the Provisions of Oxford (a written agreement between the king and magnates in England to reform the state of the realm). But the more absorbing issues of Louis's diplomacy lay in the east. He resisted papal urgings to take sides against Frederick II, believing in the equal legitimacy of empire and papacy. On the other hand, he allowed his brother Charles of Anjou to accept the crown of Sicily from the pope; for this enterprise, as for his own crusades, he allowed the papacy to tax the French clergy. His paramount foreign interest was to recover the Holy Places for Christ, a traditional ambition characteristically associated in his mind with the hope of converting the infidel: the Mongols or the emir of Tunis.

Louis IX first took the cross in 1244, upon learning that a Turkish-Egyptian coalition had driven the Christians of the Levant back to precarious coastal positions. His expedition, which was well planned and well financed, set out in 1248, only to founder in the plague-ridden floodwaters of Egypt a year and a half later. Louis himself was captured; upon his release, he spent four years in Syria in support of the Christian cause. He renewed his crusader's vow in 1267, in circumstances clouded by Angevin-Sicilian politics. Persuaded in secret by Charles, whose inordinate Mediterranean ambitions had little in common with the traditional crusade, the new expedition was diverted to Tunis. It broke up there with the king's death in 1270.

The prestige of France in Christendom lost little from these failures of Louis IX. Nor was it generally foreseen that Aquitaine and Sicily would become battlegrounds in the future.The apparent strength of his father's diplomacy deterred Philip III from changing it, even though circumstances had changed. When in 1282 the misrule of Charles of Anjou caused the Sicilians to revolt in favour of Peter III of Aragon, leading to the War of the Sicilian Vespers, a test of the Angevin policy could no longer be deferred. Charles's friend Pope Martin IV (reigned 1281–85) excommunicated the king of Aragon and offered the vacant throne to Philip for one of his sons. Because at this juncture the crown of Navarre was destined for Philip's son and successor, Philip the Fair, the whole Spanish March seemed ripe for recovery by the French. Yet the crusade against Aragon, blatantly political and impractical, came to a catastrophic end: the king himself died as his battered forces staggered out of Catalonia (October 1285). Charles of Anjou and Martin IV also died in 1285. Understandably, Philip the Fair, who had foreseen the folly of the ill-conceived attack on Aragon, no longer permitted Mediterranean concerns to dominate foreign policy. The issue over Sicily dragged on, but minor Capetian interests in the Pyrenees and in Castile were allowed to lapse.

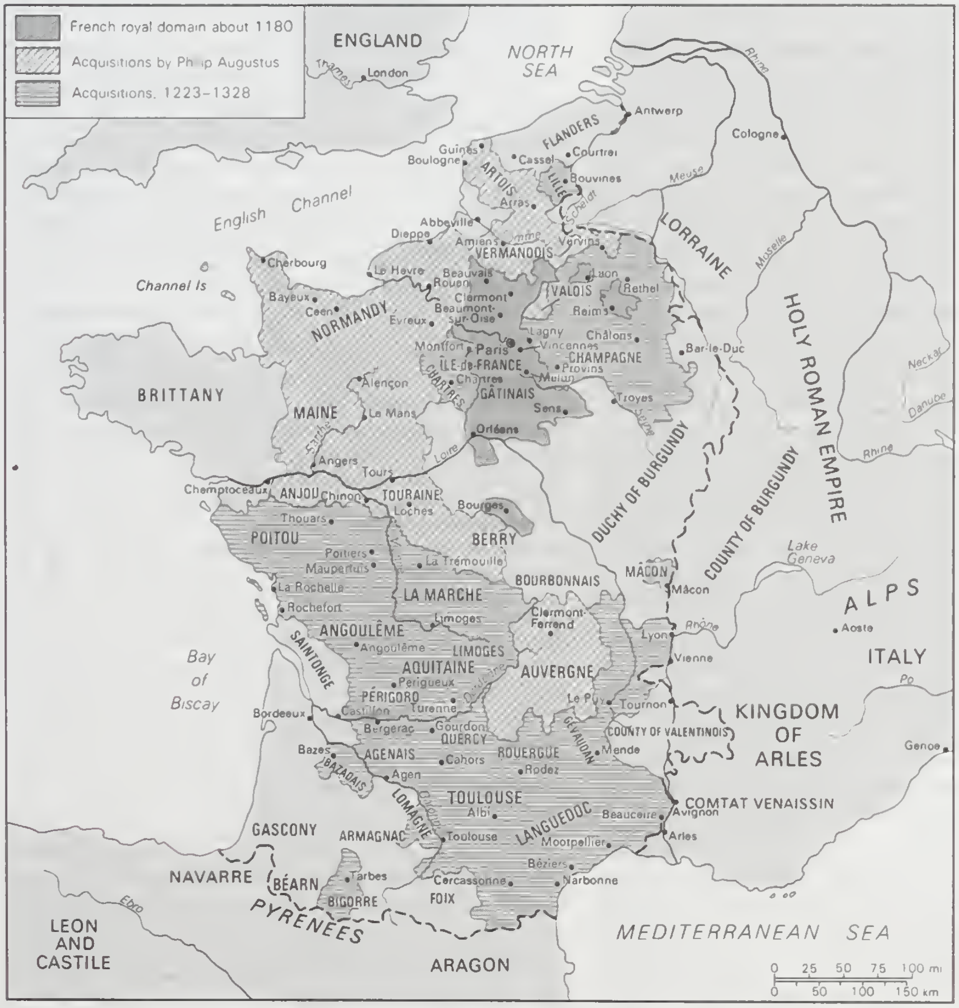
Extension of French influence to the north and east

The extension of French influence and domain toward the north and east was the result of resourceful diplomacy at the expense of the empire. Philip's interest in that direction was emphasized when his sister married the son of Albert I of Germany and when he proposed first his brother and later his son as candidates for the imperial title. But it was against the English holdings in France that Philip exercised his most aggressive and portentous diplomacy.

Questions over spheres of administrative rights in Aquitaine had been creating tensions for many years. By the Treaty of Amiens (1279) the Agenais, whose status had been left in doubt when Alphonse of Poitiers died, passed to Edward I of England, who also had unsettled claims in Quercy. Serious conflict was precipitated in 1293, when clashes between French and English seamen caused Philip the Fair to summon his vassal to Parlement. When Gascon castles occupied by the French as part of the settlement were not returned to the English on schedule, Edward renounced his homage and prepared to fight for Aquitaine. The war that ensued (1294–1303) went in favour of Philip the Fair, whose armies thrust deep into Gascony. Edward retaliated by allying with Flanders and other northern princes. His dangerous campaign, concerted with the count of Flanders in 1297, met defeat from a French force led by Robert of Artois, and during a truce from 1297 to 1303 the rival monarchs reestablished the status quo ante. Edward married Philip's sister, and a marriage was projected between Prince Edward and Philip's daughter.

A consequence of this first war was to be the chronic insubordination of Flanders. After the count's surrender

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York revision Copyright © 1964 by Barnes & Noble, Inc



The growth of the French royal domain, 1180–1328.

and imprisonment, it was left to the Flemish burghers to revolt against the French garrisons, and the French knights suffered a terrible defeat at Courtrai in July 1302. Thereafter the tide turned. But it was only in 1305 that a settlement satisfactory to the king could be reached; even then it proved impossible to win full ratification from the Flemish townsmen, whose resistance remained an invariable factor in the latent hostility between France and England.

In 1320 Philip the Fair's son, Philip V, obtained Edward II's personal homage, but friction was increasing in Gascony again. When Edward refused to do homage to Philip V's brother and successor, Charles IV, an old issue relating to French rights in Saint-Sardos (in Agenais) flamed into a war that once again went in favour of the French. By the Treaty of Paris (March 1327) France recovered Agenais and Bazadais and imposed a heavy indemnity on England, but a number of issues were left unresolved. Meanwhile, having married the emperor Henry VII's daughter, Charles was tempted to negotiate for the vacant imperial title in 1324; but nothing came of this. The last Capetians, although troubled at home, retained their international standing among neighbouring states, which were no less troubled.

Population growth

Economy, society, and culture in the 13th century. The primary social fact of this period is the continued growth of population. All indicators suggest growth—*e.g.,* expansion of old towns, founding of new villages, the rising price of land—but no exact measurements are possible. A register of hearths dating from 1328 has been estimated variously to point to a total population of 15 million to 22 million; the total was probably slightly reduced after a crest toward the end of the 13th century. By the 1280s large portions of France had enjoyed many years of relative security and prosperity, even though private warfare had not disappeared, despite royal prohibitions. Brigandage seems actually to have worsened in the south around 1200. The ravages and massacres of the Albigensian Crusade made Languedoc an insecure frontier for still another generation, though it does not appear that the Inquisition seriously disrupted urban or rural prosperity after about 1230.

The broad tendencies of social change were in keeping with political and institutional progress. The conjugal family became better recognized: Roman and especially canon law favoured its authority over the wider solidarities of clan or kin (extended family); rulers made the hearth a basis of fiscal responsibility. While mercantile practice discouraged customary restrictions on transactions, the right of repurchase by some member of the extended family held out in most areas of France. Meanwhile, a new territorial solidarity of potentially great significance was developing: that of the legal estate or class. This development was associated with the growth of towns and fostered by the consolidation of provincial custom and the fiscal impositions of count, bishop, or king. Already in 12th-century Flanders, knights (or towns) of the countryside were acting politically; the convocation of men from these incipient estates of society occurred elsewhere in France as well and more often during the 13th century. Social mobility remained considerable in spite of tendencies to restrict the ranks of urban patricians or rural nobles.

Urban prosperity. Town life continued to flourish. A few places, favoured by political and ecclesiastical as well as by economic circumstances, grew far larger than the rest. Paris could probably count more than 100,000 inhabitants by the late 13th century, possibly many more than that; some great provincial centres—*e.g.,* Toulouse, Bordeaux, Arras, Rouen—may have surpassed 25,000, but most of the older cities grew more modestly. Jewish communities, which existed almost everywhere, were especially important in the towns of Champagne and Languedoc. Immigration from the countryside probably increased as peasants sought better opportunities and independence, yet the towns remained somewhat indistinct in appearance and activity from their rural surroundings. Many urban properties had agrarian attachments, often within the walls; Paris itself was, to a surprising extent, an aggregation of expanded villages. Nevertheless, the progress of commerce, together with an important ancillary development of industry, chiefly accounts for urban prosperity in the 13th century.

Commerce and industry

The trades not only grew in volume but also became more diversified and specialized. New markets, often regional ones, arose to supplement the older centres that had developed on the basis of the long-distance exchange of relatively high-priced imperishables. Regional markets featured such agrarian staples as grains and wines as well as animals, cloth, weapons, and tools, and they facilitated the currency of foreign produce, such as glassware or spices. An increasing reliance on coinage or on monetary values may be connected with these provincial trades; sensitivity to the intrinsic values of the many French coinages was increasing everywhere toward 1200, even in the hinterlands away from main trading routes. In the late 13th century the need for money in denominations larger than the age-old penny (denarius)—primarily for use in the great commercial centres—caused Louis IX to issue the *gros tournois* (worth 12 pennies) and the gold coin (which, however, had little importance before the 14th century). A gradual long-term inflation tended to favour commercial activity.

The towns of northern France, notably in Artois, Burgundy, the Île-de-France, and especially in Champagne, prospered not only from regional exchange but also from the great overland trades connecting Normandy, England, the Baltic, and the Low Countries with the cities of Italy. The fairs of Champagne, becoming the leading entrepôt of European merchants, reached their apogee in the 13th century. Favoured by the count's privilege, the traders operated at Lagny or at Bar or—in greater numbers—at Provins and at the "warm fair" of Troyes in June; the "cold fair" of Troyes ended the yearly cycle in October. The fairs were designated as occasions for payment and repayment, contributing significantly to the progress of banking and business accounting.

Enlarged and more diversified demand encouraged urban growth and prosperity. Townsmen were eating better: in the north, at least, the per capita consumption of meat, butter, and cheese, as well as of spices, seems to have increased in the 13th century. As for wine, not only was more being drunk but the taste for *vins de qualité* became more acute, and the great regional vintages, notably that of Gascony, were established. Townspeople furnished their houses more amply than in the past (lamps, wooden chests, and draperies came into common use), and they produced more articles themselves.

The progress of industry, in fact, was a remarkable feature of the 13th century. Crafts in metal, wood, leather, and glass expanded in such large towns as Paris; clothwork—weaving, dyeing, fulling—prospered in regional centres such as Toulouse, with specialities in fine cloths concentrated in Artois and Flanders. In most places, however, the crafts remained in the shadow of commercial enterprise, in which the greater fortunes continued to be made. Artisanal associations proliferated everywhere; often termed brotherhood (*confratria, confraternitas*), they fostered new urban and suburban solidarities for charitable and ceremonial purposes as well as for the promotion of economic interests.

Urban society

Urban society became more competitive and more stratified. At Lyon, Bordeaux, and elsewhere, some fortunes were well enough established, usually from commerce, to enable their possessors to live as landlords, build stone houses, buy rural property, and aspire to titles of nobility. This patriciate, despite occasional setbacks at the hands of "new men," dominated municipal governments, acting as mayors and magistrates (*échevins*) in the north or as consuls in the south. While not altogether self-serving—they supported civic projects such as the building or decorating of churches—they were disinclined to share power or fiscal responsibility. Below them, often as their tenants or debtors, were small entrepreneurs, middlemen in trade (or between local industry and regional trade), master craftsmen, and bankers; and below all—and increasingly restive—was a swelling class of impoverished artisans, servants, vagabonds, and beggars.

Rural life. Rural life changed more gradually. The expanding markets favoured well-endowed or efficient lords or peasants who could produce a surplus of goods for sale. Such conditions were less common in the south than in the north, although they could be found in most wine-producing areas. But, while rising prices benefited producers, they contributed to certain difficulties in the countryside. Fixed revenues in coin proved an unsatisfactory alternative to payments in kind, which landlords specified when new land was put under cultivation. Moreover, needs and tastes became more expensive and tended to exceed aristocratic resources; lavish generosity continued to be an admired and practiced virtue, and costly crusades—occasionally lapsing into speculative adventures—continued to be launched. Larger lordships began to employ salaried estate managers, while in the south the division of landed fortunes among numerous heirs resulted in a multiplied and impoverished petty nobility. Many rural landlords fell into debt in the 13th century. And, as wealth and nobility became less correlated, some nobles, especially those hard pressed, sought to close ranks against the intrusion of new men or creditors. They insisted on noble birth as a condition for knighthood, while reserving the designation of "squire" (or *donzel,* in the south) for those of noble birth awaiting or postponing the expensive dubbing (*adoubement*). At the upper extreme a noble elite, the barons, achieved recognition in administration and law.

Peasant stratification

Peasant societies also became stratified. Men unable to set aside a surplus against times of famine and those who had to borrow or rent their tools or teams found it difficult to avoid dependence on other men. In some areas serfdom was renewed, or confirmed, as jurists interpreted the more stringent types of peasant obligation in the light of the revived Roman law of slavery. But here again economic and legal status did not necessarily coincide. Rich peasants who employed other men to drive their teams could be found in any village; such people as the mayor, the lord's provost, and the peasant creditor established themselves as a rural elite, whose resources insured them against calamity and opened up diverse opportunities in prospering regional economies. Where enfranchisement occurred, the lord usually received a good payment; even when servility persisted, there was a tendency to commute the arbitrary tallage into fixed common sums. New villages continued to be established, especially in the south, where many existent communities of peasants also received charters of elementary liberties—typically including the commuted tallage—in the 13th and early 14th centuries. In many cases, notably in Gascony, security, or defense, seems to have been the chief motive behind these foundations.

These conditions notwithstanding, the manor, or *seigneurie,* resisted fragmentation. The favourable market for grain and the psychological attachment of lords to their fathers' possessions preserved demesne land as the chief source of seigneurial income through the 13th century. Nor were labour services given up, although the discrepancy increased between work owed and work needed. Accordingly, lords resorted to paid seasonal labour, so that the margin between profit and loss became a more critical calculation than in the past. A new alternative was to lease the demesne to paid managers or sharecroppers, but this practice spread more slowly in France than in neighbouring countries. Whether lords had demesnes and servile tenants or not, the association between landlordship and power remained close. Tenancies or properties smaller than the old *mansus* appeared everywhere, but especially in the north, where horsepower and three-field crop rotations were making possible a more productive agriculture. The burgeoning viticultures of Burgundy and Gascony proved incompatible with traditional demesne lordship and encouraged sharecropping and peasant initiative. Innovation was less common in the uplands of the centre and south, where the manse tended to retain its identity and fiscal utility.

Religion. Whether in countryside or town, the layperson's touch with Christ became more personal and more direct. The crusade was kept alive in France by the widely shared conviction of Jewish guilt and Muslim (and Byzantine) schism. The 13th was the last medieval century in which French Jews could live securely with their Christian neighbours, but their position became progressively more difficult. The regular clergy, however, could no longer be relied upon to set standards of piety and penitence; their observance was either too relaxed or too severe to suit the new conditions. The tendency of the canonical movement of the later 12th century, moreover, had been to accentuate the distinctiveness of the secular priesthood in charge of administering to the laity and the power of holy orders. The Cistercian order, even though it continued to expand, was incapable of sustaining its ascetic impulse completely; its houses, as well as those of the older Benedictines, found their prestige and solvency alike in decline. Nor was the higher secular clergy much better situated to fulfill pastoral obligations. The bishop was by now remote from his flock, acting usually as diocesan supervisor, judge, or lord; his subordinates—the archdeacon and cathedral canons—likewise functioned primarily as administrators. Archbishops were required by the Fourth Lateran Council (1215) to hold annual synods of provincial clergy, a ruling that—although imperfectly observed—probably contributed to some strengthening of discipline.

The clergy

The critical reform was that of the parish ministry. When emphatic measures to improve the education and supervision of priests were adopted in the Fourth Lateran Council, it was already too late in France. For more than a generation, anticlericalism and doctrinal heresy had been spreading, especially in the towns and villages of the east and south. There was a suspicion that sinning priests could not be trusted to mediate God's grace effectively, and the virtue of poverty as an antidote to the worldly cupidity of a prospering society was attractive to many. The merchant Valdes (Peter Waldo), giving up his property and family (1175–76?), took it upon himself to preach in the vernacular to his fellow townsfolk of Lyon. His followers—the "Poor Men"—going sometimes so far as to administer sacraments, contrary to the canons, failed to win papal recognition. Nevertheless, the Waldensian mission, grimly tolerated by the authorities, spread to southern towns, and Valdes himself became active in the campaign against the Cathar sects of southern France. Flourishing in the hill towns and villages between Toulouse and Béziers, the Cathars held alarmingly unorthodox religious beliefs; they rejected the church as the Devil's work, organizing their own ministry.

For this challenge, the secular clergy of Languedoc was no match. To establish an effective counterministry of learned and respectable men, the pope deputed Cistercians to Languedoc; they were soon succeeded by St. Dominic of Calaruega, who spent a decade as mendicant preacher in Languedoc. In 1217, his order of preachers recognized by the bishop of Toulouse and confirmed by the pope, Dominic set out with his fellow friars to work in the wider world "by word and example."

Crusade of Innocent III in Languedoc

Meanwhile, the murder of the legate Pierre de Castelnau (1208) had stirred Innocent III to promote a crusade against the heretics of Languedoc. Led by Simon de Montfort, northern barons attacked towns in the viscounty of Béziers and later in the county of Toulouse with singular fury. Despite massacres and apostasy, the failure of the enterprise as a holy war was underscored by the establishment of a papal inquisition (1233), which stiffened and institutionalized the Dominican measures of dissuasion. The procedure, which became known in many parts of France, was usually entrusted to Dominicans; it relied on the active pursuit of suspects, secret testimony, and—in case of conviction and obstinacy—delivery of the heretic to the "secular arm" for capital punishment. Taking the war and the Inquisition together, it can be said that French heresy was largely destroyed by the end of the Capetian period.

The Franciscans as well as the Dominicans had a spectacular success. Highly organized, with provincial and international administrative institutions, both orders had houses in Paris by 1220, and their members were soon working everywhere in France. Becoming preachers and confessors, they also secured chaplaincies, inspectorships, and professorships as their initiatives in piety, probity, and learning were recognized. Conflict with the secular priesthood naturally resulted; the seculars attempted unsuccessfully to exclude the mendicants from the ministry of sacraments and inveighed against conventual endowments that seemed to contradict the friars' professions of poverty. New religious orders were soon established. The friars stimulated a more active piety among lay people, favouring charitable works and foundations, private devotions, and penitential reading.

Culture and learning. Literacy and elementary learning became more widespread. The courtly tastes of the 12th century, while not obliterated, were overtaken by a more flexible and ironic sensibility evident in vernacular ballads, fables, satires, and moralizing literature, most popular in the northern towns. The burgher or knight began to take a keen interest in the tangible world about him. The taste for clarity, proportion, and articulation, coming to mature expression in the Gothic of Amiens, Paris (Notre-Dame), and Reims, achieved dazzling finesse in the church known as the Sainte-Chapelle, built at Paris by Louis IX in 1245–48 to house the Crown of Thorns. And the taste for order is illustrated by the reorganization of masters, students, and studies as *studia generalia* (or universities); Montpellier became a leading centre of medical learning; Orléans and Toulouse, the latter founded in 1229 to prepare clerks to combat heresy, were noted for law. Paris remained preeminent among the early universities. Its famous schools became associated as the faculties of arts, canon law, medicine, and theology, gaining jurisdictional independence under papal protection by 1231.

During the same years, philosophical doctrines in conflict with Christian orthodoxy began to trouble the theologians, as translations of the metaphysical and scientific works of Aristotle and his commentators reached Paris. For a time the teaching of Aristotle was prohibited there; but by mid-century, when some of the "artists" who had been most attracted to the new philosophy were advancing to theological degrees, efforts were made to incorporate Aristotelian learning in enlarged summaries of Christian knowledge. The *Summa theologica* (1266–72) by the Italian Thomas Aquinas was the greatest synthesis of this type. Its serene power breathes no hint of the controversies in which its author was involved. St. Thomas had taken his theological degree, together with St. Bonaventure, in 1257, when the secular masters were bitterly disputing the friars' privileges within the university. In the end the Dominicans and Franciscans each retained a chair on condition of submitting to university regulations. Thomas' work, however, came under suspicion. A reaction set in against the arts faculty's increasing disposition to take a naturalistic view of all reality. The bishop's condemnations of "error" in 1270 and 1277 were so sweeping as to render even Thomas suspect.

The bishop's condemnations of "error"

Thomas' audacious brand of synthesis was to have no immediate imitators. Nevertheless, the social consequences of his newly organized learning were profound; it created new estates of professional men—lawyers, notaries, trained clerks, physicians, many of them laymen—whose rational and legalist outlook became firmly rooted in French culture.

The dogmatic condemnations of the 1270s were symptomatic. Prosperity and confidence were shaken in many ways in the late 13th century. The papacy, hitherto a support for progressive causes, found itself discredited after its fiasco in the crusade against Aragon; while the removal of the papal court to Avignon in the time of Clement V created a new centre of patronage for arts and letters, it did little to arrest the waning prestige of the church. The burdens of renewed warfare increased social tensions in the towns and depressed civic enterprise; the Jews had their assets confiscated before being expelled in 1306, and the Lombard bankers suffered like treatment in 1311. Economic indicators—while few and difficult to interpret—are generally held to suggest retardation, at least in many parts of France. The business of the fairs of Champagne was falling off by 1300, if not before, while records of Normandy reveal declining agrarian revenues in the half century after 1260. Some regions were "saturated" with people: their existent economic technology could no longer sustain growth. Probably the population was already leveling off, if not yet decreasing, when, from 1315 to 1317, crop failures and famine caused serious disruption and in some places a reversion to subsistence farming.

THE PERIOD OF THE HUNDRED YEARS' WAR

The kings and the war, 1328–1429. At the accession of the house of Valois in 1328, France was the most powerful kingdom in Europe. Its ruler could muster larger armies than his rivals elsewhere; he could tap enormous fiscal resources, including taxes authorized by sympathetic popes of French extraction; there remained only four great fiefs—the duchies of Aquitaine, Brittany, and Burgundy, and the county of Flanders—outside the direct royal domain; and the king's courts continued to press a jurisdictional supremacy that was felt everywhere in the realm. It did not follow, however, that France's superior armies would fight better than its foes or that its resources would not sometimes be dissipated or withheld. France remained a collection of traditional provinces the peoples of which believed that a king should "live off his own," while military success continued to depend on the personal leadership of dynastic rulers whose qualifications as strategists had been less refined by experience and institutional progress than their judicial or administrative competence. The history of France in the 14th century is dominated by efforts of its kings to maintain their suzerainty over the Plantagenets in Aquitaine, efforts that, despite French advantages, were long frustrated.

Philip VI. Philip VI of Valois (ruled 1328–50), grandson of Philip III, was of mature age when he became regent of France in 1328. Upon the birth of a daughter to the widow of his cousin Charles IV, the familiar issue of the succession was posed anew. It was the regent's experience, together with the circumstance that Edward III of England, grandson of Philip the Fair, was under the influence of his disreputable mother, Isabella of France, that probably disposed the council at Vincennes to recognize Philip as king (April 1328).

Philip VI and Edward III of England

Philip's reign began well. Within months he crushed a revolt of the Flemish cloth towns (Cassel, August 1328), thereby recovering the effective suzerainty over Flanders that had eluded his predecessors for a generation. And in 1329 he obtained Edward III's personal homage for the duchy of Aquitaine, an act that not only secured Philip's leadership but also nullified Edward's claim to the crown of France.

This initial success was soon undone. Jurisdictional questions in Gascony remained unsettled. In 1336 Philip VI appeared to be preparing massive support for David Bruce, the Scottish king at war with Edward; and in 1337, alleging defaults in feudal service, Philip ordered the confiscation of Aquitaine. Edward III renounced his homage and again laid claim to the crown of France, and war again was imminent. Despite the new Plantagenet pretensions, the basic causes of conflict were feudal and jurisdictional, not dynastic.

Edward proceeded deliberately and ominously. He fomented discontent among the Flemish clothworkers and then treated with the towns; in so doing he negated the count's fidelity to France; he also purchased the fidelity and service of many princes in the Rhineland and Low Countries. But, to succeed, the English needed a prompt and massive victory on French soil, something Philip VI was able to prevent. Despite Edward's naval triumph off Sluys (1340), which confirmed English control of the seas, his initial advantage was lost as his resources and allies melted away. A truce in September 1340 was extended for several years, during which time Edward intervened in a disputed succession to the duchy of Brittany, while Philip's officials increased their pressure on Gascony. In 1345 English armies counterattacked French posts on the duchy's borders; their success emboldened Edward. Landing in Normandy (July 1346) with a well-disciplined army, he captured Caen, only to be overtaken in Picardy by a much larger French army as he moved to join his Flemish allies. At Crécy (Aug. 26, 1346), despite serious disadvantages, the English forces won the first major battle of the war. Their victory, however, proved difficult to exploit; Edward moved on to capture Calais after a long siege, but

English successes

he could then only return to England with more glory than accomplishment to his credit.

Nevertheless, Philip's failures were proving costly in money and political support. In 1340–41 he had been able to raise "extraordinary" revenue through taxes on sales, salt, and hearths, despite regional protests. The continuance of sales and salt taxes in 1343 could be extracted from the Estates of Paris only in return for the restoration of a stable coinage; in the following years regional assemblies in the north proved even more obstinate. In the Estates of Paris in November 1347 the king heard ringing denunciations of his mismanagement and defeats and was fortunate to obtain new subsidies to support an invasion of England. But that prospect, like the war itself, evaporated when the Black Death struck Europe late in 1347, destroying life, fiscal resources, and resolve for several years thereafter.

Philip VI cannot be judged by his military failures alone. The royal domain was significantly enlarged by his acquisition of Dauphiné (technically an endowment for his grandson in 1343–49) and the city of Montpellier, the last (and wealthiest) Aragonese fief in Languedoc. As administrative expertise continued to progress, the services, such as Parlement and treasury, were regulated. Within the departments of the court and notably in the Chamber of Accounts, power came increasingly into the hands of royal favourites, whose rivalries were stimulated by the courtly predilections of the king. Their influence and peculation together with the familiar injustices of local government came under attack in the Estates of 1343 and 1347, which, in their conditional grants of subsidy, asserted a more nearly constitutional authority than French assemblies had yet enjoyed; the fiscal powers of the provincial Estates likewise originated during this reign.

John the Good. John II the Good (ruled 1350–64) succeeded to a weakened authority and kingdom; he was a mediocrity whose suspicions and impetuosity were ill suited to the changed circumstances. John hoped to rally baronial loyalties to himself. But he failed to reconcile Charles the Bad, king of Navarre, whose strong dynastic claim to the throne (he was the grandson of Louis X) was matched by his ambition; Charles's conspiracy—at first appeased, then too violently put down—seriously weakened John during the years (1355–56) when the English war broke out anew. When Charles sought alliance with Edward III, French diplomats abandoned full sovereignty over Aquitaine, a reversal of policy too gratuitous to hold for long; its prompt revocation, with papal support, encouraged Edward's son, the Black Prince, to undertake destructive raids through Languedoc in 1355. That November the Estates of Languedoïl, meeting at Paris, insisted on controlling the military appropriations they voted; when the Black Prince advanced from Bordeaux to Touraine in the summer of 1356, John hastened to prevent his union with rebellious Norman barons. The afmies met near Poitiers in September. Once again the French had the advantage of numbers and position, only to suffer a disastrous defeat. King John allowed himself to be taken prisoner.

The regency of the dauphin Charles

France was to experience no worse years than those of the regency, during John's captivity, of the dauphin Charles (1356–61). Unpaid or poorly disciplined armies ravaged the countrysides. The dynasts, nobles, and townspeople had new reasons to resist the monarchy. The dauphin showed no sign of adjusting to meet the crisis; the Estates-General convoked in 1356 to provide for the king's ransom demanded sweeping administrative reforms, even imposing upon the regent a council representing the Estates. Their program proved unworkable, and Charles tried to resume power on terms already rejected by the Estates. This move radicalized Étienne Marcel, provost of the Parisian merchants and leader of the urban estate. Causing the brutal murders of two of the dauphin's noble associates, Marcel only succeeded in creating an irreconcilable breach with the dauphin, who fled Paris and convoked his own assembly at Compiègne. Marcel's enthusiasm mounted as his position became more precarious; he drew strength from alliance with Charles the Bad but failed to win the Flemish towns to his cause. The climactic complication was a terrible uprising of the peasants (the Jacquerie), which broke out in Picardy in May 1358 and which antagonized Marcel's noble supporters, notably Charles the Bad, who helped to quell the disturbances. Marcel was increasingly isolated when loyalist sentiment mounted and administrative failures became evident. His assassination on July 31, 1358, not only secured the dauphin's authority but ended the burgher influence that had originated in the Estates of 1355.

Efforts to end the war

Intense efforts were then made to end the English war. Negotiations dragged past the term of truce set in 1356; when a first and too humiliating treaty was rejected by the dauphin, Edward made yet another demonstration in France (1359). At Brétigny (May 8, 1360) King John's ransom was set at three million gold crowns, while to England was assigned full sovereignty over Aquitaine (including Poitou). Two months later John arrived in Calais, where a first payment of ransom was made. In the definitive Treaty of Calais (Oct. 24, 1360), for reasons not clear, the monarchs' renunciations—Edward's claim to the crown of France, John's to sovereignty over the ceded territories—were postponed. The Black Prince, however, proceeded to take control of Aquitaine, while the regent tried with little success to extract additional money for the ransom from an exhausted country. When the Estates at Amiens (October 1363) refused to ratify an irresponsible agreement between the king's replacement hostages and Edward III, John returned to captivity in London, where he died a few months later.

Charles V. Under the former dauphin, now Charles V (1364–80), the fortunes of war were dramatically reversed. Charles had a high conception of royalty and a good political sense. While he shared the Valois taste for luxury and festivity, he reverted to the Capetian tradition of prudent diplomacy. He observed the Treaty of Calais, which helps to explain why Edward III did not press to conclude the renunciations; but he reserved his authority in Aquitaine by inserting in his coronation oath a clause prohibiting the alienation of rights attaching to the crown.

The early years of his reign were filled with baronial politics. Charles the Bad once again revolted unsuccessfully, his dynastic claim to Burgundy running afoul of the king's; the succession to Brittany was settled by arms in favour of the Anglophile Jean de Montfort (who became John IV the Valiant). Most significant for the future, Charles V obtained the heiress to Flanders for his brother Philip the Bold, to whom Burgundy had been granted in appanage. Meanwhile, companies of mercenary soldiers, many based in strongholds of central France, were paralyzing the countrysides. Charles V commissioned the Breton captain Bertrand du Guesclin to neutralize them. Between 1365 and 1369 Bertrand employed the companies in adventurous conflicts in Spain; many of the mercenaries were killed or dispersed. The Black Prince had also intervened in Spain, and his taxes and administration in Aquitaine aroused protest. In 1369 the lords of Albret and Armagnac, having refused to permit levies of subsidy in their lands, appealed to Charles V for the judgment of his court. Although Charles hesitated, his eventual decision to accept the appeals was in keeping with the letter of the Treaty of Calais and his coronation oath.

French successes

The war with England soon broke out again. Two new factors worked in favour of France. First, Charles's alliance with Henry II of Trastámara, king of Castile, cost the English their naval supremacy; a Castilian fleet destroyed English reinforcements off La Rochelle in 1372, which effectively secured the success of French operations in the west. Second, Charles abandoned the defective policy of massive engagement with the enemy. Unable to command in person, he appointed Bertrand du Guesclin constable in 1370; the latter proceeded to harry the enemy and to prey on supplies with great effectiveness. Through skirmishes and sieges, the French forces soon reconquered Guyenne and Poitou, leaving only some port towns (Calais, Cherbourg, Saint-Malo, Bordeaux) in English hands. To finance these operations, Charles continued to levy the taxes on merchandise, salt (*gabelles*), and hearths that had been intended to raise John's ransom; despite serious inequities and defaults, these taxes persisted to the end of the reign. In Languedoc they were voted, assessed, and expended by the Estates; elsewhere, by transforming the deputies first chosen by the Estates in the time of John into royal officers, Charles created a fiscal administration independent of popular control. His military success owed much to the improved regulation of armed forces and defenses. Ordinances provided for the inspection and repair of fortifications, the encouragement of archery, a more dependable discipline, pay for fighting men, and even the establishment of a navy.

The last years of the reign brought disappointments. Truces were arranged; but, as there could be no more talk of ceding French sovereignty over Aquitaine, there could be no assurance of peace. More serious, the papal–French alliance collapsed. Charles V, unable to prevent Pope Gregory XI from returning to Rome in 1376, chose to support the candidacy of Robert of Geneva against the Italian Urban VI in 1378, but only Scotland and Naples followed the French lead. A schismatic pope could no longer help France much; rival popes could hardly promote peace between their political supporters. Although he had reestablished the political unity of France, Charles V left an uncertain future.

Charles VI. Charles VI (ruled 1380–1422) was a minor when he succeeded his father. His uncles, each possessed of the ambition and resources to pursue independent policies, assumed control of the government. Duke Louis of Anjou soon removed himself from influence by seeking the throne of Naples; Jean, Duke de Berry, received the lieutenancy of Languedoc, by then virtually an appanage; and it was left to Duke Philip the Bold of Burgundy to set the young king's policy. He imposed his own cause upon the king in his policy toward Flanders (whose ruler, Count Louis II, was Philip's father-in-law). An uprising by the workers of Ghent, spreading to other towns, was met by royal force that won a crushing victory at Roosebeke in 1382. The young king returned in triumph to deal forcefully with restive populations at Paris and Rouen and in Languedoc. The provostship of the merchants was suppressed at Paris, bringing that municipality under direct royal control.

In 1388 Charles VI assumed full authority himself. He recalled his father's exiled advisers, the Marmousets, who undertook to reform the royal administration in keeping with the practice of Charles V. But the country was again wearying of taxation. The annual levies of Charles V had been discontinued in 1380 but then reestablished—helping to cause the urban unrest already mentioned—and were being dissipated blatantly in royal and princely extravagance. In 1392 the king lost his sanity, a shocking event that aroused popular solicitude for the crown. His recurrent lapses into insanity, however, played into the hands of his uncles. Philip the Bold again dominated the council. Fortunately for France, England was incapable of renewing the war. The Duke of Burgundy planned an invasion of England in 1386, but after major preparations in Flanders it came to nothing. A series of truces, beginning in 1388, was followed by a reconciliation between Richard II of England and Charles VI in 1396, when the truce was extended for 28 years. Meanwhile, French nobles were reviving the crusade, imagining a reunited West following their lead; John the Fearless' defeat at Nicopolis in 1396 was the most famous of several enterprises. To restore unity in the church, the masters of the University of Paris began to speak out vigorously; the conciliar theory (according to which the church was to be governed by an ecumenical council), which finally prevailed to end the schism, owed much to them.

Insanity of Charles VI

When conflict with England was renewed in the 15th century, circumstances had changed. Henry IV of England was committed to the recovery of English rights in France; moreover, in a civil war between Louis, Duke d'Orléans, and John the Fearless (duke of Burgundy since 1404) over control of the king, both parties sought English support. And, when John caused Orléans to be assassinated in Paris (Nov. 23, 1407), the popular horror magnified the conflict. John exploited the situation by pressing for reforms; his rival's cause was taken up by Bernard VII of Armagnac, whose daughter married Orléans's son. But John's alliance with the turbulent Parisians was no more secure than the temper of the angriest burghers; a major ordinance for administrative reform (1413) collapsed in a riot of the butchers, and in the ensuing reaction the Armagnac faction regained control of Paris. John's dangerous response was to encourage the new king of England, Henry V, to claim the French throne for himself. Henry's invasion of 1415, reminiscent of the campaign ending at Crécy, had the same result—at Agincourt the French suffered yet another major defeat, after which, characteristically, the English withdrew—but the civil war in France enabled Henry V to exploit his strength, as Edward III had not been able to. In 1418 the Burgundian party recovered control of Paris, and the dauphin Charles embarked on a long exile in Armagnac company.

But John's duplicity was limitless; while meeting with the dauphin to betray the English, John was himself assassinated (1419). His successor, Philip the Good, renewed the alliance with Henry V. By the Treaty of Troyes (1420) the deranged Charles VI was induced to set aside the dauphin's right of succession in favour of Henry V, who married Charles VI's daughter. The ancient dream of a dynastic union between France and England seemed to be realized; and, when Henry and Charles died within weeks of each other in 1422, the infant Henry VI became king in both lands.

Charles VII. Charles VI's son, Charles VII (ruled 1422–61), for his part, did not fail to claim his inheritance, though he had no proper coronation. Residing at Bourges, which his adversaries pretended was the extent of his realm, he in fact retained the fidelity of the greater part of France, including Berry, Poitou, Lyonnais, Auvergne, and Languedoc. For a time the Valois cause suffered from the ineptness of its leader and from his advisers and retainers, who prospered from the unresolved conflict. Incapable himself of military leadership, Charles put his hope in reconciliation with Philip of Burgundy, a diplomacy that thoroughly discomfited King Henry's regent, the Duke of Bedford. Nevertheless, French prestige collapsed with the abasement of the monarchy; Charles VII appears to have doubted his own legitimacy, and disorders spread again.

Joan of Arc

Then Joan of Arc appeared. Stirred by the popular memory of traditional French kingship, she found her way from her peasant home at Domrémy (on the border of Champagne and Bar) to Chinon, where she confronted Charles with her astonishing inspiration: her "voices" proclaimed a divine commission to aid the king. In April 1429 she entered Orléans, long besieged, rallying the garrison to effective sorties that soon caused the English to lift the siege. Other victories followed, in which Joan's influence was manifest, although probably exaggerated in tradition. On her insistence that only consecration at Reims could make a true king, chosen by God (a view doubtless supported by the chancellor Regnault, archbishop of Reims), it was decided to advance boldly across the Île-de-France to Reims. Charles was anointed there on July 17, 1429.

Recovery and reunification, 1429–83. The coronation of Charles VII was the last pivotal event of the Hundred Years' War. From Reims the king's army moved on triumphantly, winning capitulations from Laon, Soissons, and many lesser places and even threatening Paris before disbanding. The popular devotion to monarchy that had produced Joan was undermining English positions almost everywhere in France; the urgent necessity to discredit her explains the callous efficiency of the inquisition to which she was subjected upon being captured by the Burgundians and turned over to the English in 1430. Condemned of heresy, confessed under duress, boldly relapsed, she was burned at the stake in Rouen on May 30, 1431.

Charles and his party made no move through ecclesiastical channels to save Joan. They then proceeded deliberately to make peace with Burgundy. In the Treaty of Arras (Sept. 21, 1435), Philip the Good bargained strongly; confirmed in the possession of domains ceded by the English, he also obtained Charles' humiliating disavowal of the murder of the duke's father, John the Fearless. The act, however damaging to the royal vanity, set Charles free from political obligation to the Armagnacs; the factional king now became the supreme king of France. Within a year, English support collapsed in the Île-de-France, and royal soldiers entered Paris. The Truce of Tours (1444) provided for a marriage between Henry VI and the niece of Queen Mary of France; extensions of the truce gave Charles time to strengthen his military resources. War flared again in 1449, when England intervened against a duke of Brittany who had done homage to Charles VII. In 1449–50 a vigorous campaign resulted in the French conquest of Normandy, and in 1451 most of Guyenne fell to the French.

The end of the Hundred Year's War

When the English lost the minor battle of Castillon in 1453, the Hundred Years' War was over. That fact was not altogether clear to contemporaries, for no treaty was concluded and skirmishes were to recur for many years to come. But only Calais, enclosed in the Burgundian domains, remained of English possessions in France. Charles VII issued medals to commemorate his soldiers, and he ordered a review of Joan of Arc's trial, which resulted in a verdict of rehabilitation in 1456.

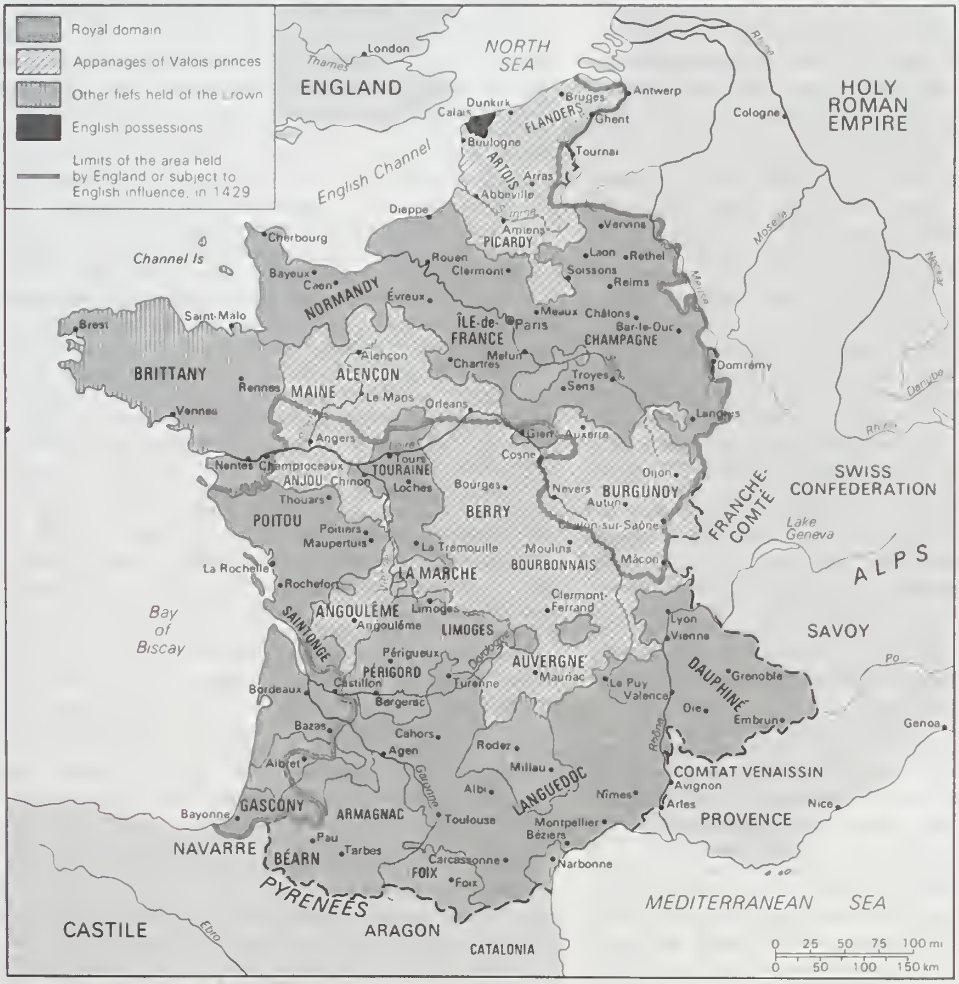
Governmental reforms. As hostilities were waning (1435–49), Charles VII presided over a major reorganization of government. Tested by adversity and strengthened by fortune, he had grown in political competence. The principal administrative services—chancery, Parlement, accounts—were reestablished at Paris. The replacement of Burgundian sympathizers, notably in Parlement, seems to have been accomplished with moderation and tact; in local offices no purges were necessary. But it quickly became evident that the reunited country was now too large and its officials too numerous to get along very well with a government as centralized as Parisian bureaucrats preferred.

Remedial legislation was consistent with tendencies long apparent. Revenues from the domain were collected in the treasury, the work of which Charles VII reorganized in four regional offices. Extraordinary revenues had been administered since the 1350s in districts (*élections*), the numbers of which had vastly increased since the time of Charles V. The *élections* were now subordinated to four regional *généralités*, corresponding to the offices of treasury. The old Chamber of Accounts had lost parts of its jurisdiction to more specialized courts in 1390, of which the Cour des Aides (board of excise) had provincial divisions set up at Toulouse in 1439 and at Rouen in 1450. A provincial *parlement* was definitively established at Toulouse in 1443, and there were to be others at Grenoble and Bordeaux. With all these changes the conciliar structure of government survived; policy continued to be made by the king in concert with favourites, whose numbers no reforms had delimited. The proliferation of lesser offices, many filled by lawyers, created a new stratum of gentlemen who enjoyed the king's privilege.

Revenues

While the reform of offices did nothing to obliterate the older distinction between ordinary and extraordinary revenue, the work of Charles VII effectively belied the notion that the monarchy should subsist on its domain alone. That the king as lord could no longer pay his officers and soldiers was apparent to almost everyone. Early in his career Charles had resorted to the Estates to raise *aides* and *tailles* (as the old levies on sales and hearths were now called), but after convocations in the 1430s he continued these taxes through annual ordinances no longer sanctioned by the Estates. Moreover, the preparation of annual budgets for ordinary and extraordinary revenues gave way in 1450 to a single "general statement" of finance, which, being related to demonstrable necessities, effectively institutionalized taxation in France. As the Middle Ages ended, France comprised a central core of *élections*, where local Estates, when they met at all, had little to do with fiscal matters, and a surrounding belt of "lands of Estates" (*e.g.*, Languedoc, Brittany, Normandy, and Burgundy), where custom continued to allow for consent or for the administration of taxes. Having originated in times of

From W Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York revision Copyright © 1964 by Barnes & Noble, Inc



France in 1453.

fiscal demands thought uncustomary and excessive, representative institutions could not generally survive once the royal impositions, from very repetition, had ceased to seem arbitrary; even where Estates persisted, their votes were more like approval than sovereign consent.

The new standing army

Military reforms. The fiscal reorganization facilitated equally significant military reforms. The Peace of Arras, rather than pacifying France, had only thrown the people once again to the mercies of disbanded mercenaries and brigands. In 1439 an ordinance made the recruitment of military companies the king's monopoly and provided for uniform strength in contingents, supervision, and pay. Following the Truce of Tours in 1444, no general demobilization occurred; instead the best of the larger units were reconstituted as "companies of the king's ordinance," which were standing units of cavalry well selected and well equipped; they served as local guardians of peace at local expense. With the creation of the "free archers" (1448), a militia of foot soldiers, the new standing army was complete. Making use of a newly effective artillery, its companies firmly in the king's control, supported by the people in money and spirit, France rid itself of brigands and Englishmen alike.

Regrowth of the French monarchy. Thus the monarchy recovered much of the authority it had lost during the early stages of the Hundred Years' War. Although its influence in Burgundy and Flanders, now united in a formidable dynastic association, had declined, its definitive recovery of Aquitaine consolidated a direct domain, again extensive enough to free the Valois from anxiety about landed resources. It had exploited not only a widespread distaste for the destructive self-interest of barons and warlords but also an incipient nationalism, which, besides reviving the "religion of monarchy," put new stresses on the foreignness of Englishmen. How renewed power and Gallicanism went together was demonstrated in the Pragmatic Sanction of Bourges (1438), by which papal benefices and revenues from France were severely curtailed and the royal influence in the French church strengthened. Nevertheless, the survival of powerful dynasts and provincial interests, as a legacy of the war and the fertility of the royal house, represented a counterpoise to the crown that Philip the Fair had never known. And, with the son of Charles VII, the monarchy was to be tested yet again.

The tyranny of Louis XI

Louis XI (1461–83) was shamelessly impatient for his father's death. It must be said of this strange man that he had worthy policies to pursue: the securing of the royal domain against Burgundy, Orléans, and Brittany, among others, and the promotion of commerce and industry within national boundaries. His foreign policy was less consistent, ranging from the cautious in Italy to the chimerical in Spain; yet it was at the expense of Aragon that he regained title to Roussillon and Cerdagne. His methods rather than his ends were what made the reign of this ambitious, nervous, and capricious ruler so turbulent. No French king had ever imposed himself so totally and so tyrannically as Louis XI. Forgetful of past loyalties, he was betrayed as often as he himself betrayed others. Toward the clergy as toward his officials he could be brutal and vindictive. He antagonized the nobles by revoking the Valois pensions and ceremonial and by promoting the independence of seigneurial towns. As for the royal towns, Louis respected their constitutions only so far as was consistent with royal supervision and the payment of heavy taxes; he tolerated the resurgence of urban oligarchies. Fiscal pressures in support of the army, government, and diplomacy mounted fearfully.

Politics, under Louis XI, replaced administration as the foremost preoccupation of the realm. Arbitrary and hasty measures against the dynasts aroused the formidable League of the Public Weal, which, in 1465, appealed to the people against misgovernment and proposed a regency of the princes supported by the three estates. Louis, in turn, as on later occasions, used assemblies and proclamations to divide the princes. But the settlement of October 1465 was a grave setback for the king, whose brother Charles gained title to Normandy while Charles the Bold, soon to inherit Burgundy, acquired strategic counties and towns in Artois. To the undoing of this treaty Louis devoted great energy. Fomenting strife between Brittany and Normandy, he soon recovered the latter and isolated the former. Deaths among his rivals in Gascony enabled him to secure successions, as in Armagnac, more divided and less hostile. Increasingly, Louis's tortuous diplomacy fastened on Burgundy. The king succeeded in reconciling the Swiss cantons with Austria to form a coalition with France and the Rhenish cities; this coalition invaded Burgundy and defeated and killed Charles the Bold at Nancy (Jan. 5, 1477). While the legal reversion of Burgundy to the crown could not be given practical effect, Louis did recover Artois. Moreover, even as he enjoyed this decisive triumph over his most dangerous rival, the entire Angevin inheritance (Anjou, Provence, and Mediterranean claims) devolved to the crown upon the death of René of Anjou in 1480. Through accident and design and the inability of the princes to collaborate effectively, Louis had succeeded in countering the threat of a princely constitution and had considerably extended the royal domain.

Economy, society, and culture in the 14th and 15th centuries. The long war, fought almost entirely in France, benefited few but the captains and peculators; it injured almost everyone. Even the best disciplined companies lived off the land, so that French peasants and defeated townsfolk in effect paid the expenses of both sides; and undisciplined mercenary bands were a wearisome scourge in times of truce after the middle of the 14th century.

Economic distress. But the war did not alone cause economic distress. Even before it broke out, bad weather and commercial dislocations, together with overpopulation in some areas, resulted in worse famines and more frequent ones than in the past. But what most terribly damaged life and security was sickness.

The Black Death

The Black Death, carried on shipboard from the Levant, reached Provence in 1347, ravaged most of France in 1348, and faded out only in 1350. Nothing worked to check the disease in populations without immunity—neither bonfires to disinfect the air, nor collective demonstrations of penitence in northern towns, nor persecutions of Jews or friars. The mortality was staggering—the French chronicler Jean Froissart's estimate that the first wave carried off a third of the population was perhaps not far wrong; evidence shows that rural areas were no less afflicted than towns. And there were recurrent outbreaks of plague in later years. These afflictions and related factors were responsible for a general decline of population. Toulouse seems to have lost half of its population, which fell from 40,000 to 20,000; the population of Normandy is estimated to have declined by two-thirds between 1300 and 1450. The trend was probably reversed, and perhaps strongly so, in the second quarter of the 15th century, although little is yet known about this.

The hard times affected classes and regions in different ways, degrees, and rhythms. Some places almost escaped the ravages that afflicted others repeatedly. In the countrysides, especially—save for the greatest personages—those who had most to lose suffered most. Whether for landlords or rich peasants, surpluses became harder to obtain or preserve; to many lesser lords the dangerous fortunes of war probably seemed an attractive alternative to declining yields in money or produce. Standards of living, as measured in diets or furnishings, declined. Onerous obligations and services tended to disappear as shortages of rural labour made themselves felt; the transition from servile to rental tenures was largely completed in the 15th century. Peasant uprisings, such as the Jacquerie in the relatively prosperous Île-de-France and the Tuchins in Languedoc, are too poorly documented to be well understood—both betrayed desperation born of recurrent taxation and were associated with the expression of egalitarian ideas; the Jacquerie coincided with a weakened grain market and may have been hastened by efforts of lords to enforce labour services and payments after the Black Death. The manor survived, but little remained of its human identity in the 15th century. Even minor lords lived away from their peasant tenants, protected them poorly if at all, relied on salaried managers to collect payments that, in some cases, had lost all social justification; lordship had degenerated into an unsentimental economic practice.

The cities. Urban society was also troubled. The walled town stood out ever more starkly against the countryside as siege warfare intimidated or destroyed the suburbs that had been built in a less anxious day. Royal taxation, often inequitably administered, exacerbated old tensions in the towns; fiscal policy or the regulation of wages or supplies was largely at issue in the uprisings of Flemish towns (1323–28), at Paris (1357–58, 1380–82), and at Rouen (1382). Communes continued to be revoked in the 14th century, although the kings as a rule were less interested in governing the towns than in securing their resources and fidelity. The concentration of trades and crafts in guilds became more complete and more exclusive.

Decline of commercial centres in the 13th century

Some leading commercial centres of the 13th century suffered as new trade routes developed in the empire and by sea and as textile manufactures and money markets—the latter suffering from unstable coinages—became more dispersed. The fairs of Champagne declined rapidly after 1310. Only a few capitals, such as Avignon, Bordeaux, and Paris, prospered; and even they were hard hit by plague. Nor did the French merchant or manufacturer progress competitively; his work often unspecialized, his bookkeeping old-fashioned, his tastes simple, he typically looked forward to securing his future by the purchase of land.

Form of Christian piety

The church. The organized church, despite losses from war and plague, continued to be better endowed economically than morally. The popes of Avignon were less distant and—save perhaps to their French relatives, merchants, and artists—less admirable than the reformer popes of the past; as the prestige of their schismatic successors plummeted, the higher clergy were confirmed in their incipient Gallicanism (a movement advocating administrative independence from papal control). While organized heresy had almost disappeared, reforms intended to strengthen the parish priesthood languished. Jurisdictional disputes continued to rage between mendicants and seculars and between bishops and canons or archdeacons. Christian piety even more than in the past sought encouragement in mystical or individual devotions or readings and in collective observances of the Holy Spirit or the Virgin or the patron saints of the trades that promoted elementary solidarity and charity in the towns; such confraternities were not always welcome to ecclesiastical authorities, whose deportment or jurisdiction they sometimes challenged, whether directly or merely by example. The popular religion of saints—more particularly of the Virgin and the Pietà—and fear of demons worked more deeply into the collective imagination, becoming very evident in the 15th century. Associated with intensified anxieties about sin and damnation, these experiences thrived in times of recurrent and inscrutable disaster.

Culture and art. Cultural circles remained strongly oriented to aristocratic values and the past. With the accession of the Valois came a high nobility, distinguished by lavish and exclusive conceits. When John II formed the Order of the Star (1351), an institution imitated by the great lords for their clientages, chivalry stood incorporated as the most distinguished of religious confraternities. The ideal of the crusade remained strong, notably among princes of the fleur-de-lis, who dominated the public life of Valois France to the point of eclipsing the monarch; beneath them many noble families disappeared, while new ones emerged among the captains, lawyers, and patricians. Froissart spun out chronicles of the war at once detailed and grand, full of the frivolous ceremonial that marked the aristocratic life of his day. Tapestries created for courtly patrons idealized a life of enticing gardens, tournaments, and the hunt. Paintings as well as tapestries decorated the walls of chambers that were smaller and more elegant than the cavernous halls of earlier centuries. The rayonnant Gothic of the Île-de-France remained in favour through the 14th century, inspiring the chapel built by Charles V at Vincennes, while the decorative arts of furnishings and manuscripts exploited the Gothic tendencies to articulation and grace. The evocation of the classical past became less fantastic and more heroic in the humanist circles of Pierre Bersuire and Petrarch; their interests helped to attract copyists and artists to the papal court of Avignon. The *Book of Hours* (the most popular private devotional work of the later Middle Ages) might become "very rich," as in the case of a sumptuous manuscript undertaken for Jean, Duke de Berry (*c.* 1410); more typically it was a pocketbook for general use by the literate, whose numbers continued to increase.

Music

Stimulated by the commissions of Charles V, the chasm between learned and vernacular cultures narrowed: Raoul de Presles translated St. Augustine; Nicole Oresme translated Aristotle. Music resounded in old forms (ballad, virelay) even while becoming more articulate or flamboyant; Guillaume de Machaut (d. 1377), the great musician-poet of the mid-14th century, composed the first polyphonic mass as well as many motets and secular lyrics. Time and space came to be better represented and measured, as evidenced by the first attempts to render perspective in art and in the erection of public clocks at Paris and Caen.

Toward 1400 Paris regained cultural leadership as a result of a new synthetic (or international) style in painting and of the initiatives of the university masters in ecclesiastical politics and theology. The efflorescence, however, was soon destroyed in the civil wars, to be succeeded later in the 15th century by more provincial activities. Universities (like *parlements*) proliferated at the expense of Paris, which became the preserve of an antiquated and pedantic theology. Painters, architects, and writers regrouped under princely patrons or even under bourgeois ones, flourishing in postwar trade (Jacques Coeur's palace at Bourges exemplifies the flamboyantly decorated solidity of late-medieval taste in France). A new style in painting, as in architecture, characterized by vigour and an enlarged scale, contrasted with the more traditional style in Burgundy, where the dukes were building on a grand and continuous past. Italianate humanism, together with the new philology, stirred in France only in the latter third of the 15th century. (T.N.B.)

France, 1490–1715

FRANCE IN THE 16TH CENTURY

When Charles VIII (1483–98) led the French invasion of Italy in 1494, he initiated a series of Italian wars that were to last until the Peace of Cateau-Cambrésis in 1559. These wars were not especially successful for the French, but they corresponded to the contemporary view of the obligations of kingship. They also had their effects upon the development of the French state; in particular, they threatened to alter not only the military and administrative structure of the monarchy but even its traditional role.

Military and financial organization. The French kings of the early 16th century could look back with satisfaction at the virtual expulsion of the English from French soil in the course of the preceding century. This success offered a shining precedent for further military sallies, this time against the growing power of the Habsburgs. In 1445 the first steps had been taken to fashion a royal French army out of the ill-disciplined mercenary bands upon which French kings had traditionally relied. It was a small force—no more than 8,000 men—but it was a beginning. The role of the nobility in the army was strong, for the art of war was still considered a noble pursuit par excellence. The core of Charles's army that marched into Italy, the *compagnies d'ordonnance,* known collectively as the *gendarmerie,* consisted of noble volunteers. The infantry, however, was made up of nonnobles, and by the middle of the 16th century there were more than 30,000 infantrymen to a mere 5,000 noble horsemen. As this infantry force grew in number, its organization changed. Legions, organized on a strict provincial basis (the Breton Legion, the Norman Legion, etc.), gave way to the regimental system, based on large units under a single command. This latter organization appeared during the Wars of Religion of the 16th century and survived until the time of Louis XIV. Of great significance, too, was the involvement at the heart of the royal army of the provincial governors as commanders of the *gendarmerie.* Yet such reorganization did not immediately reduce the army to a pliant tool of the crown. Not until late in the 17th century could the royal army be considered fairly under the king's control. Until then, notably during the Wars of Religion and the outbreaks of the Fronde (1648–53; see below), the loyalty of the commanders and the devotion of the troops were conspicuously inadequate. In the later part of the 17th century, the reforms of the army by Michel Le Tellier and his son the Marquis de Louvois provided Louis XIV with a formidable weapon.

The problem of royal finance

The growth of a large royal army, however, was only one effect of the increased level of military activity. The financial administration of the country also underwent a drastic reorganization, which had far-reaching economic and social consequences. The king, despite his ambitions, possessed neither the resources nor the administrative machinery to maintain a large army. The medieval idea that the king should live off the revenue of his own domain persisted into the 18th century and helps to explain the formal distinction made until the reign of Francis I (1515–47) between ordinary and extraordinary finance—*i.e.,* between revenue emanating from the king's patrimonial rights and taxes raised throughout the kingdom. By the reign of Francis I, the king, even in times of peace, was unable to make do with his ordinary revenue from rents and seigneurial dues. In 1523 Francis established a new central treasury, the Trésor de l'Épargne, into which all his revenues, ordinary and extraordinary, were to be deposited. In 1542 he set up 16 financial and administrative divisions, the *généralités,* appointing in each a collector general with the responsibility for the collection of all royal revenues within his area. In 1551 Henry II added a treasurer general; from 1577 the *bureaux des finances,* new supervisory bodies composed of a collector general and a number of treasurers, made their appearance in each *généralité.*

The taille

The actual collecting of taxes, moreover, was increasingly handed over to tax farmers. The more efficient methods of collection by tax farmers enabled the crown to gather a larger proportion of its revenue than previously but did not solve the problem of royal finance. Even the extraordinary taxes, now added to the crown's ordinary revenue, notably the taille (a direct tax levied on all but the nobility and the clergy), custom duties, and the purchase tax on wine, fish, meat, and especially salt (the gabelle), were not adequate resources for Renaissance princes whose chief glory lay in the expensive art of war. The taille, the only direct tax, which weighed most heavily upon the underprivileged classes, went up from about 4.5 million livres under Louis XI (1461–83) to 55 million under Jules Cardinal Mazarin in the mid-17th century.

Successive monarchs were forced, therefore, to seek additional revenue. This was no simple matter because French kings traditionally could not tax their subjects without their consent. Indeed, there were many areas of the country where the taille itself could not be collected and where the king was dependent upon local agreements. The early Valois kings had negotiated with the Estates-General or with the provincial Estates for their extra money; but in the middle of the 15th century, when the Hundred Years' War with England was reaching a successful conclusion, Charles VII was able to strike a bargain with the Estates. In return for a reduction in overall taxation, he began to raise money to support the army without having to seek the Estates' approval. In some areas of central France the provincial assemblies ceded their right to approve taxation and disappeared altogether. These provinces were known as the *pays d'élection.* But, in those provinces where the provincial Estates survived (the *pays d'état*), the right to vote the amount of royal taxation also survived. During the Italian wars, meetings of the Estates became more frequent as the king's financial demands became more strident, and, though the Estates never felt themselves able to refuse to provide money, they retained the right to provide less than the monarch asked for. The king continued to rely upon the support of the provincial assemblies to provide extra revenue long after 1614, when the cumbersome Estates-General ceased to play a role in opposing financial resources for the crown.

The sale of offices

The growth of a professional bureaucracy. But the king also found another means of filling his exchequer that had nothing to do with traditional methods: he began to sell offices on a large scale. Venality, or the sale of offices, was not novel in early 16th-century France; traces of the practice can be found in the 13th century. But it was Francis I who opened the floodgates. The number of judges proliferated. In the Parlement of Paris alone the king created two new chambers, each containing 20 members and a further score of judges. In 1552 Henry II established a new kind of court, the *présidial,* whose jurisdiction lay between the *parlement* and the bailiwick. Each of the 65 new courts had a complement of nine judges; this brought in a sizable revenue but appears to have made little difference to the efficiency of the judicial system. Nor were judicial offices the only ones put up for sale; it was also possible to purchase financial offices, such as those of treasurer general, treasurer, or the immediately inferior *élu.* It has been estimated that during the 16th century some 50,000 offices were sold by the crown.

The partial rationalization of the financial system produced an increasing number of professional advisers, who formed the embryo of a bureaucratic elite. In the course of the 16th century, as specialization grew apace, the king's council became a much more complex institution. The Conseil d'État, with its various subdivisions, formed the hub of royal government. Its members were drawn from a variety of backgrounds. The king's immediate family expected to be consulted, as did great officers of the crown, such as the chancellor, the constable, and the admiral. Also included in the council were the great territorial magnates, members of powerful aristocratic families, and the country's leading prelates. There were also masters of requests (*maîtres des requêtes*), lawyers whose expertise was invaluable when the council sat in a judicial capacity. But in the council the professional element that assumed the greatest significance in the course of the 16th and 17th centuries were the holders of the office of secretary of state. In the early years of the 14th century, royal secretaries had already acquired the right to sign documents on the king's authority. From this stage, granted the stability of the crown, the development of the office from a position of subordinate but considerable importance to one of complete indispensability was predictable. Henry II gave four of his secretaries the official title of *secrétaire d'état,* and in 1561 they became full members of the royal council. Closely associated with them and destined to overshadow them in importance in the first half of the 17th century were the superintendents of finance, formally established in 1564, though exercising an already well-established function. Their responsibility was to control and safeguard royal finances and especially to prepare annual budgets containing estimates of revenue and expenditure for the following year. They also played a leading part in assessing the amount to be levied each year from the taille and in deciding upon the imposition of new taxes. Below the superintendents but also in the royal council in the 16th century were the intendants of finance. Originally masters of requests, they became a separate group specializing in the increasingly complex task of advising the sovereign in financial matters. In time, their role outstripped in prestige that of the other masters of requests who counseled the king.

The new administrators

There thus grew up a more specialized class of administrators, close to the crown, whose expertise rather than birth was the key to their influence; the sale of office allowed wealthy families to establish a firm base for later political and social advancement. In addition, the needy crown was perfectly prepared to sell titles of nobility as well as offices and, in return for a cash payment, to allow both nobility and office to become hereditary. Although this advancement of new men within the government might suggest a social readjustment of considerable proportions, in fact the element of continuity was more important than might at first appear. Even though it is true that some of the ancient noble families and the king's own relatives found it increasingly difficult to fulfill their old advisory roles, the new men were not rejecting the established order but rather were being absorbed into it. The king's counselors, whatever their former background, became leading noblemen by virtue of their high office: service to the crown was what mattered, and elevation to the office depended on the king's choice. It was not the first time that a new wave of royal servants had begun to overtake established advisers; in the 13th century the *magistri* had ousted the great barons and prelates from the Curia Regis without effecting a social revolution. What took place in the 16th and 17th centuries was another turn of the social wheel by which new men seized the opportunity to pursue those dignities and honours held by men who were themselves descendants of new men.

Jacques Lefèvre d'Étaples

The Reformation. The professional class that grew up in the 16th century, however, was different in one respect from those that had gone before: it represented a predominantly secular culture—the product of Renaissance humanism. Nevertheless, in the second half of the 16th century the Reformation was to embroil France in a series of religious wars that were to pose a serious threat to the power of the king and his government. Lutheran works first appeared in Paris in 1519; in 1521 Francis I, who was on the point of war with Emperor Charles V and King Henry VIII of England and who wanted to demonstrate his orthodoxy, forbade their publication. Yet, interest in the new faith continued to grow, especially in the humanist circle of Jacques Lefèvre d'Étaples. Lefèvre, who had in 1512 published an edition of the letters of St. Paul with a commentary that anticipated Martin Luther in its assertion of the doctrine of justification by faith, became the leader of a small group of moderate but orthodox reformers in the tradition of the great Dutch humanist Desiderius Erasmus. This group included Guillaume Briçonnet, the bishop of Meaux; the mystic Gérard Roussel; and Margaret of Angoulême, the king's own sister. Although this circle was dispersed in 1525, Lutheranism had already established itself, especially in such trading centres as Lyon where it found support among the poorer classes. The progress of the Reformation in France depended on the crown's attitude; although Francis for political reasons had initially shown hostility, his feelings were far from clear. He was favourably disposed toward Lefèvre and toward orthodox reform in general, though he naturally feared those extreme movements that threatened social upheaval. In addition, Francis I saw political advantages in establishing good relations with the Lutheran German princes. On the other hand, unlike them, he had no great incentive to assert his independence from Rome because the Gallican church already enjoyed a large measure of autonomy. In 1516 the Concordat of Bologna had given the king effective control over the church in France.

The influence of Calvinism

In 1534, however, royal policy changed radically. Anti-Catholic placards began to appear in Paris and other major French towns, provoking a bitter Catholic reaction and a series of persecuting edicts. French Protestantism itself had changed, reinforced from the mid-1530s by the spread among the poorer classes of Languedoc and the seaboard towns of Normandy and Brittany of the ideas of John Calvin, a French exile in Geneva. Henry II (1547–59) pursued his father's harsh policies, setting up a special court (the *chambre ardente*) to deal with heresy and issuing further repressive edicts, such as that of Écouen in 1559. Yet the infusion of French Calvinism, or Huguenotism, into the French Reformation stiffened the Protestant opposition. Protestant pastors, trained in Geneva, infiltrated into the country; by 1562 there were some 2,000 highly organized Calvinist churches in France. This spectacular spread of Calvinism persuaded the queen mother, Catherine de Médicis, who was ruling in the name of her young son, Charles IX (1560–74), to abandon the repressive religious policy of Francis I and Henry II in the name of political good sense. Guided by the moderate chancellor Michel de L'Hospital, Catherine summoned the French clergy to the Colloquy of Poissy (1561), at which an unsuccessful attempt was made to effect a religious compromise with the Huguenots; in the following year she issued the Edict of January, which allowed the Calvinists a degree of toleration. These signs of favour to the Protestants brought a violent reaction from the noble House of Guise, the champions of Roman Catholicism in France. The first civil war began with the massacre by the partisans of François, Duke de Guise, of a Huguenot congregation at Vassy (March 1562).

The Wars of Religion. Guise forces occupied Paris and took control of the royal family, while the Huguenots rose in the provinces, and their two commanders, Louis I de Bourbon, Prince de Condé, and Admiral Gaspard de Coligny, established headquarters at Orléans. The deaths of the opposing leaders—the Protestant Antoine de Bourbon, King Consort of Navarre, and the Catholic marshal Jacques d'Albon, Seigneur de Saint-André—and the capture of Condé caused both sides to seek peace. After the Battle of Dreux (December 1562) the war drew to a close, despite the assassination of the Duke de Guise by a Protestant fanatic. A compromise was reached at the Peace of Amboise in March 1563: liberty of conscience was granted to the Huguenots, but the celebration of religious services was confined to the households of the nobility and to a limited number of towns.

The second war was precipitated by Huguenot fears of an international Catholic plot. Condé and Coligny were persuaded to attempt a coup to capture Catherine and Charles IX at Meaux in September 1567 and to seek military aid from the Protestant Palatinate. In the following brief war, the Catholic constable Anne de Montmorency was killed at the Battle of Saint-Denis (November 1567); the Peace of Longjumeau (March 1568) signaled another effort at compromise. This peace, however, proved little more than a truce; a third war soon broke out in September 1568. In an attempt to restore their authority, Catherine and King Charles dismissed L'Hospital in September and restored the Guise faction to favour. The edicts of pacification were rescinded; Calvinist preachers faced expulsion from France, and plans were made to seize Condé and Coligny. The former was killed at the Battle of Jarnac (1569), and the Huguenots were again defeated in that year at Moncontour. But the Catholic side failed to consolidate its successes, and yet another compromise was arranged at the Peace of Saint-Germain in August 1570.

Massacre of St. Bartholomew's Day

Coligny subsequently regained the king's favour but not the queen mother's, and he remained an object of hatred with the Guises. In 1572 he was murdered. At the same time some 3,000 Huguenots, gathered together in Paris to celebrate the marriage of Marguerite de Valois (later Margaret of France) to Condé's nephew, Henry III of Navarre, were massacred on the eve of the feast day of St. Bartholomew. This notorious episode was the signal for the fifth civil war, which ended in 1576 with the Peace of Monsieur, allowing the Huguenots freedom of worship outside Paris. Renewed fighting broke out in 1577 between Catholic and Protestant noblemen, who defied the king—now Henry III (1574–89)—in his attempt to assert royal authority. The Huguenots were defeated and forced by the Peace of Bergerac (1577) to accept further limitations upon their freedom. An uneasy peace followed until 1584, when, upon the death of François, Duke d'Anjou, the Huguenot leader Henry of Navarre became the heir to the throne. This new situation produced the War of the Three Henrys (1585–89) during which the Guise faction—led by Henri, Duke de Guise—sought to have Navarre excluded from the succession. In a welter of intrigue and murder, first the Duke de Guise and his brother Louis, Cardinal de Guise (December 1588), and then Henry III himself (August 1589) were assassinated. Henry of Navarre thus assumed the throne as Henry IV (1589–1610), the first king of France from the house of Bourbon (a branch of the house of Capet); he had to survive five more years of civil war and embrace Roman Catholicism before his position was secure. In its final stages, the war became a struggle against Spanish forces intervening on behalf of Isabella Clara Eugénie, the daughter of Philip II of Spain and Elizabeth of Valois, who also laid claim to the French throne. The Peace of Vervins (1598), by which Spain recognized Henry IV's title as king and the Edict of Nantes of the same year, granting substantial religious toleration to the Huguenots, ended the Wars of Religion.

The religious wars and the monarchy. This succession of civil disturbances brought the French state close to disintegration and posed a threat to the crown that would not be matched again until 1789. The key factor in producing this situation lay in the nature of Calvinism, which provided both a rallying point for a wide cross section of opposition and the organization necessary to make that opposition effective. Each Huguenot community created its own administrative structure to provide a tight disciplinary framework, through which the community could ensure its spiritual and material independence. The new creed attracted several elements in French society: small artisans, shopkeepers, and the urban unemployed, who were suffering in particular from steeply rising prices; many rich townspeople and professional men who thought that material advancement would be easier to procure as Calvinists; and, after the Treaty of Cateau-Cambrésis in 1559, many nobles, especially the poorer ones who had lost with the peace their best hope of wealth and status.

The social composition of Calvinists

The adherence of large numbers of the nobility had two important effects upon the movement in France: it caused many peasants to join the new creed in imitation of their noble seigneurs, thus swelling the overall number and widening its social composition, and it brought a new military element into the Calvinist communities. Under the leadership of the nobility, secret religious meetings were transformed into mass public demonstrations against which the king's forces were impotent. Such demonstrations sometimes involved upward of 20,000 people. Similarly, the administrative structure that was so important in aiding the survival of the proscribed faith was transformed into a military organization. This organization was ultimately headed by Louis, Prince de Condé, who assumed the title of protector general of the churches of France, thus putting all the prestige of the house of Bourbon behind the Huguenot cause. By doing so, he added a new dimension to the age-old opposition of the mighty feudal subject to the crown: that opposition was now backed by a tightly knit military organization based on the Huguenot communities, by the financial contributions of wealthy bankers and businessmen, and by the dedicated religious zeal of the faithful—inspired by the example of Geneva.

At a time when the threat to the crown had never been greater, the monarchy itself presented a sorry spectacle. The struggle for political power at the centre of government after Henry II's death between the families of Guise, Bourbon, and Montmorency; the vacillating policy of Catherine de Médicis; and, most important of all, the ineptitude of three successive rulers—Francis II (1559-60), Charles IX, and Henry III—meant that local government officials were never confident of their authority in seeking to curb the growing threat of Huguenotism. Indeed, the chief opposition to Protestantism came not from the crown but from the Catholic Holy League. Because of the government's inability to control the situation, local Catholic unions or leagues began to appear in the 1560s, headed by nobles and prelates. In 1576, after the Peace of Monsieur with its concessions to the Huguenots, these local leagues were fused into a national Catholic Holy League. The league was headed by the Guise family and looked for material aid to Philip II of Spain. Its chief aims were the defeat of Protestantism in France and the restoration of ancient feudal rights and privileges. It sought, like the Protestants, to attract mass support; its clandestine organization was built around the house of Guise rather than the monarchy, from which it was increasingly alienated. In 1577 Henry III tried to nullify the league's influence, first by putting himself at its head and then by dissolving it altogether. This maneuver met with some success.

But in 1585, when the death of Henry's brother, who had succeeded him as Duke d'Anjou the previous year, made a Protestant succession to the throne much more likely, a second and far more revolutionary organization appeared. This second movement was centred in Paris among middle-class professional men and members of the clergy and soon spread among the Parisian artisans, guilds, and public officials. Finally, through the intervention of the Duke de Guise, it inspired the reappearance throughout the country of the old Catholic Holy League of 1576, though now in a much more extreme and threatening form. The king himself, who was considered far too tolerant toward the Huguenots, was an object of attack. In town after town, royalist officials were replaced by members of the league. In Paris, the mob was systematically aroused; in 1588, in the famous "Day of the Barricades," Henry III was driven from his own capital. After the murder of Guise at the end of that year, the league came out in open revolt against the crown. Towns renounced their royal allegiances and set up revolutionary governments. In Paris, however, where the league was most highly organized, a central committee called the Sixteen was established. It set up a Committee of Public Safety and conducted a reign of terror in a manner similar to the much more famous one during the revolution of 200 years later.

The Catholic Holy League in Paris

Paradoxically, this genuinely democratic and revolutionary element in the Holy League paved the way for the triumph of the Protestant leader Henry of Navarre, later Henry IV. The aristocratic members of the league took fright at the direction in which the extreme elements in the movement were proceeding. Their fears reached a climax in 1591, when the Sixteen arrested and executed three magistrates of the Parlement of Paris. The growing split in the ranks of the members of the league, combined with Henry's well-timed conversion to Roman Catholicism, enabled him to seize the initiative and enter Paris, almost unopposed, in 1594. But the threat to the monarchy and therefore to the whole French state had been of a new and fundamental kind, and the strong position that Henry IV achieved by the time of his death is that much more remarkable. Part of his success lay in the unwillingness of his great subjects to contemplate a social and political upheaval that would displace them as well as the king from their positions of power and prestige.

Political ideology. The religious wars also engendered a luxuriant growth of political ideas that in the end provided a strong theoretical basis for the reassertion of royal authority.

A strong element in Calvin's teaching was the importance of passive obedience to secular authority—an idea that became impossible for the Huguenots to support after the Massacre of St. Bartholomew's Day. They began instead to advocate the right to attack the king if he would not guarantee them toleration. The most important Huguenot contribution in this change was the anonymous pamphlet *Vindiciae contra tyrannos* (1579), which raised fundamental questions about the prince's power and the rights of his subjects. The pamphlet advanced the idea of a twofold contract: the first contract, between God and ruler on the one hand and the ruler and his subjects on the other, recognized the belief that the king ruled under the aegis of Divine Providence; the second contract, between the king and the people, obliged the king to govern justly and the people to obey him so long as he did so. It followed from the argument in the *Vindiciae* that subjects had the right to rebel if the prince disobeyed the laws of God or refused to govern his people justly. This twofold contract was not intended to be a license for private and personal rebellion but was interpreted as justifying the corporate opposition of whole towns and provinces.

A second element in the realm of political ideas, deeply opposed to the contractual theory of the Huguenots, was that of the Jesuit supporters of Ultramontanism. The Ultramontanists feared that a strong national monarchy would mean the subordination of the church to its authority and the diminution of papal authority. They feared the triumph of both Huguenotism and Gallicanism in France. Their most effective controversialist was the Italian prelate Robert Bellarmine, whose *Disputationes* (3 vol., published 1586–93) and *De potestate summi pontificis in rebus temporalibus* (1610) gave definite form to the theory of papal supremacy. By no means were all members of the league supporters of Bellarmine, though their extreme Catholicism made many of them sympathetic to his ideas. The definitive Gallican reply came in 1594 with Pierre Pithou's *Les Libertés de l'église gallicane,* which reiterated the basic tenets of Gallican doctrine: that the pope had no temporal authority in France and no more spiritual power than that bestowed on him by such conciliar decisions as the monarchy chose to recognize.

The Ultramontanists

The growing support for Gallican opinion was a reflection of the emergence of the Politique Party after the Massacre of St. Bartholomew's Day. In the opinion of this moderate Catholic group, toleration should be granted to the Huguenots for the sake of peace and national unity. The Politiques were the spiritual heirs of the chancellor

L'Hospital and represented an attitude of mind rather than an organized movement. Under the pressure of political events this group became convinced of the need to support a strong monarchy that could resist both Ultramontane and Huguenot excesses and the divisive influence of noble factions. They therefore increasingly identified themselves with the Gallican position. The Huguenots, too, were not slow to see the advantages for themselves of this new attitude, and the ideas of the *Vindiciae* gave way to the theory of passive obedience. The wheel had turned full circle.

Divine-right theories

With this emphasis upon passive obedience emerged the theory of the divine right of kings. The first written statement of the theory in France is contained in the works of Pierre de Belloy, especially his *De l'autorité du roi* (1588). He asserted that the monarchy was created by God and that the king was responsible to God alone. Any rebellion against the ruler, therefore, was a rebellion against the Almighty. The essential premise of the divine-right idea is that the right to command obedience cannot be bestowed by man; only God can grant such authority. God therefore chooses the king, and there can be no contractual relationship between the king and his people; to rebel even against an unjust ruler is to challenge God's choice. If the king breaks his contract with God, then he is answerable to God alone. On the wave of such ideas Henry of Navarre became king of a united France, supported by Huguenots and moderate Politique Catholics alike. The universalist doctrine of Bellarmine gave way to the national one of Pithou as the country closed ranks against Spain, the common enemy.

One other concept emerged about this time that helped to set the seal on Henry's authority: the idea of sovereignty, as expounded by Jean Bodin. In his *Six Livres de la République* (1576) Bodin argued that the political bond that made every man subject to one sovereign power overrode religious differences. Bodin provided the link divine right did not allow between the king and his people; divine right was concerned with the source of the ruler's power, sovereignty with its exercise. The needs of the political situation forced Bodin to give his sovereign virtually unlimited authority, though he insisted—as was traditionally the case in France—that the ruler should respect the sanctity of the natural law, of the fundamental laws of the kingdom, of property, and of the family. In 1614, on the occasion of the last meeting of the Estates-General before the Revolution, the Third Estate sought to have it made a fundamental law of the realm that under no pretext whatever was it permissible to disobey the king. This effort gives some indication of the extent to which the ideas of divine right and sovereignty had provided a firm theoretical base for the reestablishment of monarchical power after the dangerous years of civil war.

FRANCE IN THE EARLY 17TH CENTURY

Henry IV. The restoration of royal authority was not, of course, simply a matter of adjusting theories of kingship; there was a clear practical reason for Henry's success. The country had tottered on the brink of disintegration for three decades. By the time of Henry's succession, it was generally recognized that only a strong personality, independent of faction, could guarantee the unity of the state, even though unity meant religious toleration for the Protestant minority. By the Edict of Nantes (April 13, 1598) Henry guaranteed the Huguenots freedom of conscience and the right to practice their religion publicly in certain prescribed areas of the country. As a surety against attack, the Huguenots were granted a number of fortresses, some of them, such as La Rochelle and Montpellier, extremely formidable. Huguenots were made eligible to hold the same offices as Roman Catholics and to attend the same schools and universities. Finally, to ensure impartial justice for them, the Edict established in the Parlement of Paris—the supreme judicial court under the king—a new chamber, the Chambre de l'Édit, containing a number of Protestant magistrates who would judge all cases involving Huguenots. Although the problem of religion was not finally settled by the Edict of Nantes, Henry did succeed in effecting an extended truce during which he could apply himself to the task of restoring the royal position.

Sully's financial reforms

The chief need of the monarchy was to improve the financial situation, parlous since the days of Henry II's wars and aggravated by the subsequent internecine conflict. Henry was fortunate in this connection to have the services of Maximilien de Béthune, Duke de Sully, who was admitted to the king's financial council in 1596. Sully at once embarked upon a series of provincial tours, enforcing the repayment of royal debts, thereby increasing the king's revenues. He also provided the first real statements of government finances in many years; by 1598 he had become the effective head of the royal financial machine as well as a trusted member of the king's inner Cabinet. He held a variety of offices: superintendent of finances, grand master of artillery, superintendent of buildings, governor of the Bastille, and others. But it was in the field of finance that he made his greatest contribution to the welfare of the state. Sully was not an original financial thinker. He undertook no sweeping changes, contenting himself with making the existing system work, for example, by shifting the emphasis from direct to indirect taxation. He succeeded in building up both an annual surplus and substantial reserves.

The only measure Sully championed that might be described as novel and far-reaching was the introduction in 1604 of a new tax, the *paulette,* named after the financier Charles Paulet, which enabled officeholders to assure the heritability of their offices by paying one-sixtieth of the purchase price each year. The *paulette* was intended to increase royal revenues, though it had considerable political implications too, in effect making government offices practically hereditary. Politically, the *paulette* was to increase the independence of a wide range of royal officials, thereby limiting royal absolutism and strengthening the possibility of disinterested state service. In addition, Sully did much to reorganize fortifications and to rebuild roads and bridges after the devastation of the religious wars. In transportation his greatest work was the Briare Canal project to join the Seine and Loire rivers—the first such scheme in France—completed under Louis XIII.

Sully, however, favoured a much more cautious domestic policy overall than did his sovereign; because Sully disliked merchants and manufacturers, he opposed many of the king's economic ventures. Henry IV believed in direct state intervention, and he took steps to fix wages and to prohibit strikes and illegal combinations of workmen. Henry's policies bore fruit especially in the textile industries, where the production of luxury silk goods and woolen and linen cloth greatly increased. Henry also took the initiative in making commercial treaties with Spain and England, thereby increasing the volume of French trade and stimulating the export of grain, cattle, and wine. Yet his efforts were not entirely successful, not least because merchants were more concerned with buying land and office (and thereby status) than with plowing back their profits into further industrial development. Though the country did assume a more prosperous air under Henry IV, that change was chiefly because of the domestic and foreign calm that followed the Peace of Vervins.

Henry's foreign policy

Even after Spain's agreement in 1598 to the restoration of the territorial position as it had existed in 1559, Henry was not free of international complications. But he was able to prevent them from once more dividing his kingdom. He did have to counter a conspiracy led by one of his own marshals, Charles de Gontaut, Duke de Biron, who plotted with the king of Spain and almost succeeded in raising southwest France in revolt. Henry, however, had Biron arrested and executed in 1602; this strong action against an old friend and powerful enemy had the effect of subduing the political rising and strengthening Henry's own authority. In central government Henry gave increasing power to Sully at the expense of the rest of his council, while in the provinces the responsibilities of the intendant, an official first regularly employed during the reign of Henry III, were widened to include the supervision of potentially dissident groups. The intendants also represented the crown at meetings of provincial estates, enforced royal laws, and advised the king on a variety of local problems—fiscal, administrative, and military. When Henry IV was assassinated by François Ravaillac, a Catholic fanatic, in

May 1610, he had gone a long way toward restoring the monarchy to a position of authority similar to that held by Francis I and Henry II and had reunified a state greatly threatened at his accession from both within and without.

The rise of Richelieu

Louis XIII. Henry's reign was followed by the regency of his widow, Marie de Médicis, who ruled on behalf of their young son Louis XIII (1610–43). Once more the security of the country was threatened as factions disputed around the throne. The work of Henry IV seemed likely to be undone. Crown and country, however, were rescued by probably the greatest minister of the whole Bourbon dynasty—Armand-Jean du Plessis, Cardinal de Richelieu. Richelieu first came to the attention of the government in 1614, when he was chosen to present the final address of the clergy at the meeting of the Estates-General. His eloquence and political expertise on this occasion won him the notice of Marie de Médicis, who later appointed him her secretary. By 1616 Richelieu was secretary of state for war and foreign affairs. His career, however, received a check in the following year when a palace revolution overthrew the regency of the queen mother, exiling her to Blois. Richelieu was banished first to Luçon and subsequently to Avignon (1618). He began the climb back to power by negotiating the Treaty of Angoulême (1619), which reconciled Louis XIII to his mother. After the death in 1621 of Louis's favourite, Charles d'Albert, Duke de Luynes, Richelieu regained effective power; he became a cardinal in 1622 and in April 1624 gained access to Louis XIII's council. On the disgrace in 1624 of the superintendent of finance, Charles de La Vieuville, he became Louis's principal minister—a position which he maintained until his death some 18 years later.

Richelieu proved an indefatigable servant of the French crown, intent on securing absolute obedience to the monarchy and on raising its international prestige. The first objective required him to crush a number of revolts of the nobles, the first of which, in 1626, involved the king's younger brother and heir, Gaston de France, Duke d'Orléans. Louis acted ruthlessly, and one of the conspirators, Henri de Talleyrand, Count de Chalais, was executed. Then, in 1630, came the celebrated "Day of Dupes" when the king's life was despaired of and the queen mother, now allied with Gaston and the keeper of the seals, Michel de Marillac, prepared to move against Richelieu. The king, however, recovered and chose to support Richelieu against the wishes of his mother, his wife, and his confessor. Finally, at the very end of his life, the cardinal had to overcome another conspiracy headed by the young royal favourite, Henri Coiffier de Ruzé, Marquis de Cinq-Mars, in which Gaston was once more implicated. Through all these crises Richelieu retained the king's support, for it was in Louis's interests, too, that such intrigues should be firmly dealt with.

In the course of strengthening royal absolutism, Richelieu also came into conflict with the Huguenots. He believed that their right under the Edict of Nantes to maintain armed fortresses weakened the king's position at home and abroad. Protestant rebellions in 1625 and 1627 persuaded the cardinal of the need for a direct confrontation. The major Huguenot citadel of La Rochelle was attacked by royal troops in 1627 and, despite attempts by the English to assist the Protestants, fell in the following year. Another royal army marched into Languedoc, where the Huguenot forces were concentrated, and quickly overcame them. The Peace of Alais (1629) left the Huguenots free to enjoy religious and civil liberties, but they lost the military power that had made them a threat to the government. They were never to pose that sort of threat again, and little more would be heard of them until Louis XIV decided to repeal Henry IV's Edict of Nantes.

Richelieu also took a great interest in economic matters. To promote economic self-sufficiency, he encouraged the manufacture of tapestry, glass, silk, linen, and woolen cloth. He gave privileges to companies that established colonies in America, Africa, and the West Indies. To protect trading and colonial interests, he created a navy, which by 1642 had 63 oceangoing vessels.

Richelieu's anti-Habsburg policy

On the basis of these policies, Richelieu was able to pursue an increasingly successful foreign policy. His first aim was the security of France, which he hoped to achieve through the occupation of key points on the country's frontiers lying along imperial and Spanish territories. He thus involved France in the War of the Mantuan Succession (1628–31) in northern Italy. Through diplomatic means he worked for the dismissal of Albrecht Wenzel von Wallenstein, the brilliant general fighting on the side of the emperor Ferdinand II, whose forces were threatening to destroy the Protestant princes of Germany in the Thirty Years' War. To undermine the power of the Habsburgs, he prolonged this conflict, negotiating with the United Provinces; with Gustav II Adolf of Sweden, with whom he concluded the subsidy Treaty of Bärwalde in 1631, agreeing to pay the Swedish king one million livres per year to continue the war; with Gustav's successor, Count Axel Oxenstierna; and with Bernhard, Duke of Saxe-Weimar. Eventually, in 1635, Richelieu committed France to direct conflict with the Habsburgs; and before his death he had savoured the triumph of having French arms in the Spanish Netherlands, Lorraine, Alsace, and Roussillon.

The career of Richelieu bears something of a contradictory aspect. He undoubtedly added to the earlier success of Henry IV and Sully in overcoming the threat of anarchy and disorder that was the legacy of the late 16th century. Indeed, his contemporary reputation was one of supreme ruthlessness and arbitrariness in the application of power. Yet he was never more than the king's creature, incapable of pursuing a course of action of which Louis disapproved, always vulnerable to the loss of royal favour and support. He was ambitious, but he recognized that his desire for power could best be satisified within the confines of dutiful royal service. Richelieu was no innovator: he devised neither new administrative procedures nor novel methods of taxation to secure the king's authority. Indeed, the power of the great financiers grew with the government's need for additional war revenue, posing a different threat to royal absolutism. Richelieu's unique contribution lay in the single-minded devotion he gave to the task of increasing royal authority at home and abroad. He also succeeded in accumulating a vast personal fortune as a result of his years in power. Richelieu died in 1642, and Louis XIII died the following year. France was once again ruled by a regent, the queen mother, Anne of Austria. But the task of governing the country fell increasingly into the hands of another cardinal, Jules Mazarin.

Opposition to Mazarin

The Frondes. The years of Louis XIV's minority were dominated by the Frondes, a series of civil disturbances that lasted from 1648 to 1653. The government's financial difficulties were once more at the root of the trouble. In the first few years of the regency a variety of expedients were tried to raise additional revenue for the war with Spain. There was about these expedients an air of arbitrariness and compulsion that antagonized a wide cross section of Parisian society, notably the Parlement of Paris, and the animosity was heightened by Mazarin's use of intendants in the localities to cut across traditional legal hierarchies. Although most of the disputes were on the face of it concerned with financial exactions, below the surface an older constitutional argument was developing, as Mazarin followed Richelieu in attempting to dictate from the centre in the interests of the state. The climax came when the government failed to renew the *paulette* for the members of the provincial *parlements* and for some of the chief legal officeholders in the capital, in the Cour des Aides, the Chamber of Accounts, and the Great Council. This decision was not a gratuitous rebuff to these magistrates but yet another attempt to gain additional revenue, this time by offering a renewal of the *paulette* in lieu of four years' salary.

At this point, the first Fronde (the Fronde of the Parlement) began with the outraged magistrates of the three courts concerned joining with the Parlement of Paris to demand redress. Their demands included the abolition of the office of intendant; a reduction in the level of the taille; and the restoration of normal judicial procedure in registering financial edicts in the Parlement. The regent and Mazarin at first took a conciliatory attitude, but each side gradually moved to more committed and extreme positions, and civil disturbances in Paris exacerbated an already delicate situation. The magistrates increasingly aimed their fire at Mazarin, for he, like Richelieu before him, seemed to be taking over the king's authority and using it in uncharted and illegal areas. The magistrates, however, were not revolutionaries, and the state of disorder in the capital frightened them. That fact, allied with fears of a Spanish invasion (for the war was still continuing with Spain despite the Peace of Westphalia in 1648), persuaded them, in 1649, to make the Peace of Rueil with the government, the terms of which were for the most part favourable to the magistrates' original demands. At this stage the second civil war broke out, the Fronde of the Princes, headed by the Great Condé. The second Fronde was a pale reflection of the feudal reaction during the Wars of Religion; and, although Condé succeeded in gaining control of Paris, he did not acquire the support of the Parlement except briefly and under duress. In October 1652 Condé fled to Spain, and Louis XIV reentered his capital in triumph.

The Fronde of the Princes

Neither Fronde posed the grievous threat to the very basis of the state that had existed in the previous century. Mazarin was the chief object of enmity, and that fact itself helps to explain the less serious nature of the threat. What was at issue was not the king's authority per se but the manner in which it had been exercised since Richelieu's time, in a less personal and therefore seemingly more arbitrary fashion.

After the Frondes, Mazarin continued to play a key role in government as chief adviser to the young king, whose respect and affection he had long possessed. His career ended on a high note with a successful conclusion of the war with Spain negotiated by the Treaty of the Pyrenees (1659). According to its terms, France gained Roussillon and Cerdagne in the south and Artois and a number of border towns in the north; and the Rhine became France's frontier in the east. By the treaty, too, Louis XIV was betrothed to the infanta Marie-Thérèse, the elder daughter of Philip IV of Spain. It was by any reckoning a triumphant peace, though it sowed the seeds of future European conflict over the issue of the Spanish succession (see below). When Mazarin died in 1661, Louis was confident enough to take up the reins of government without recourse to another first minister.

The European crisis of the 17th century

Before examining the most famous reign in French history, one further aspect of the previous decades requires scrutiny—the economic and social situation to which some historians have attached considerable importance, especially as an element in the so-called general European crisis of the 17th century (a series of upheavals—among them the Frondes and the English Civil War—that affected Europe in the middle of this century). A Soviet historian, Boris Porshnev, first drew particular attention to the series of local uprisings taking place in France in the first half of the 17th century. He noted that all of them contained an element of desperation as serious financial hardship brought about by exorbitant tax demands and crop failures forced the participants into action. Great fluctuations in prices and outbreaks of famine further accentuated the misery. From these observations he drew ideological conclusions that have been widely disputed. Porshnev maintained that the state was the instrument by which the dominant economic class exploited the underprivileged and that revolt was a sign of the popular resistance of the exploited masses. On the other hand, Roland Mousnier, perhaps the greatest French authority on the period, has pointed to evidence that suggests that these revolts were not always spontaneous: they were provoked in 1632 by the municipal authorities at Lyon; in 1636 by the nobility of Périgord; and in 1641 by Louis, Count de Soissons, a prince of the blood. In 1643 the conduct of bishops and magistrates in Languedoc was seen by one royal official as likely to encourage sedition there. From these examples and others like them, Mousnier concludes that the nobility was seeking to reverse the process by which the king's officials were extending the crown's authority at their expense, taxing peasants to such a degree as to limit the amount the peasants could afford to pay in seigneurial dues, insinuating the king's justice between themselves and their peasants, reducing their prestige in the local community, and trampling upon old rights and privileges. Mousnier asserts that the nobility, far from representing the oppressive state, was itself involved in a struggle against increasing state interference. The opposition provoked by the work of Richelieu and Mazarin fits well into Mousnier's view and gives it support. It remains true, however, that most popular revolts began as protests against fiscal demands that were intolerable to people already suffering grievous economic hardship and that they were often provoked by a particular incident, such as the appearance of a new tax collector or a new system of tax collection.

There was little sign of the revolutionary attitude that had characterized aspects of the 16th-century Wars of Religion. On the contrary, there were positive indications of continuing loyalty to the crown, with such rebel slogans as "Vive le roi sans la gabelle" or "Vive le roi sans la taille." Nor was the other great bastion of the establishment, the church, attacked. The substantial tax of the tenth continued to be paid to the church without complaint. During the Frondes, neither the nobles nor the magistrates represented a revolutionary political element, and even the rioters on the Parisian streets were simply desperately poor. There was no social revolution either under Henry IV, Richelieu, or Mazarin despite the appearance of increasing numbers of hereditary officeholders, aided in their dynastic ambitions by the device of the *paulette*. The elevation to noble rank of new families was a cyclical process in France that would continue through and beyond the reign of Louis XIV.

THE AGE OF LOUIS XIV

Throughout his long reign Louis XIV (1643–1715) never lost the hold over his people he had assumed at the beginning. He worked hard to project his authority in the splendid setting of Versailles, to depict it in his arrogant motto: "Nec pluribus impar" ("None his equal"), and in his sun emblem. He buttressed his authority with the divine-right doctrines elaborated by Bishop Jacques-Bénigne Bossuet and proclaimed it across Europe by force of arms. Yet he made surprisingly few institutional or administrative changes in the structure of government. Like Richelieu, Louis used the system that he had inherited and adapted it to suit his own personality and outlook. This practice may be seen first in his attitude to the machinery of central government.

Louis's inner council

The development of central government. Louis's inner council was on the model of the royal council in Richelieu's days, a High Council (Conseil d'en Haut) consisting of only three or four members and excluding the king's own relatives. Members of this council were known as ministers, but they held no formal right to the title and ceased to be minister if the king chose not to summon them. The first of these great men were Michel Le Tellier, Hugues de Lionne, and Nicolas Fouquet; but the latter was disgraced within a year, and by 1665 his place had been taken by Mazarin's former secretary, Jean-Baptiste Colbert. These three men dominated the government in the early years of Louis's personal reign, but always, as with Richelieu and Louis XIII, under the watchful and jealous eye of the king. Le Tellier had been secretary of state for military affairs under Mazarin's regime, and his greatest contribution under Louis was to reorganize the army along lines that were hardly changed until after 1789. He created a royal army that wore the king's uniform; it was commanded by his officers and was ultimately responsible to the sovereign. It was a standing army of hitherto undreamed-of size, reaching 400,000 men in times of war and requiring close regulation in matters of discipline, training, recruitment, supply, and overall organization. The success of Le Tellier and of his son Louvois, who succeeded him, goes far to explain the dominance of French arms in Europe during Louis's reign.

Lionne, the expert in foreign affairs, had been the chief French negotiator at the Peace of the Pyrenees. His effective influence with Louis is difficult to gauge; he certainly was not the sole source of advice in foreign affairs. Lionne remains a more elusive personality than his colleagues, though there can be no doubt of his importance. It should be remembered that all important matters of state were reviewed at the High Council; and the king's ministers were expected to give advice and opinions on all that was discussed, not simply on matters in the area of their particular expertise.

Colbert's financial reforms

Colbert, however, remains the best-known of these intimate counselors. Of the 17 ministers summoned by Louis XIV to the High Council during his reign, 5 were members of the Colbert family. In 1664, Colbert was appointed superintendent of the king's buildings; in 1665, controller general of finances; in 1669, secretary of state for the navy. His capacity for work and his grasp of detail were remarkable; but he was not an original, much less a revolutionary, thinker. His chief contribution to the king's finances, like Sully's, was to make the machinery more efficient, not to substitute any new mechanisms. Colbert's first achievement was to present the king with a monthly statement of the financial situation, though his annual estimates for the following year never persuaded Louis of the need for economies if his mind was set in other directions. Yet, within 10 years of taking office, Colbert, mainly by tightening up on the tax-collecting administration and by rationalizing the gathering of indirect taxes, did succeed in producing a surplus. He turned a large part of central and northern France into a free-trade area and gave the responsibility for collecting all indirect taxes there to a new syndicate of tax farmers called the Farmers-General. Under Colbert, the total sum levied from indirect taxation rose from 36 million livres to 62 million.

In his industrial policy Colbert believed that France needed to produce for itself those manufactured goods that it was having to import. To achieve this mercantilist goal, derived from, among other sources, the ideas of Richelieu, Colbert was willing to invoke a variety of improvisations: direct subsidies, exemptions from the taille, monopoly grants, controls exercised through town guilds. Skilled foreign workmen were persuaded to settle in France and to pass on their skills to native artisans; protective tariffs were imposed. The famous tapestry works of the Gobelins family was made a state enterprise, and France became largely self-sufficient in the production of woolen cloth. Colbert also had some success in other industries, such as sugar refining, plate-glass making, and the production of silk, naval stores, and armaments. The overall results of his hard work, however, were disappointing. France underwent no industrial revolution during the reign of Louis XIV.

Growth of the navy

Much more successful were Colbert's efforts at fostering the growth of the navy. He reorganized the recruitment system on a rotary basis, whereby seamen served in the royal navy for six months in every three years. He refurbished the hospitals in each of the major ports; rebuilt the arsenals at Toulon and Rochefort; and increased the size of the navy from about 25 ships in 1661 to 144 in 1677. He also established schools of marine engineering, hydrography, and cartography. His interest in reestablishing French sea power was, in part, to challenge the commercial supremacy of the Dutch. He encouraged the building of the French mercantile marine and established a number of overseas trading companies, in particular the East India and Levant companies, neither of which had much success. He also attempted to protect French colonial interests in the West Indies and Canada.

The king's council

Besides the High Council, the king's council also met for somewhat less vital matters under a variety of different guises. The Council for Dispatches (or, more loosely, the Council for the Interior; Conseil des Dépêches) had particular responsibility for home affairs, including the activities of the intendants; the Royal Council for Finances (Conseil Royal des Finances) supervised important matters affecting financial aspects of the king's domain lands. These two councils, like the High Council, were presided over by the king in person. But the royal council also met under three further titles to deal with judicial and administrative matters and not in the king's presence. The Privy Council (Conseil Privé) judged disputes between individuals or bodies and dispensed the king's supreme and final judgments. The State Council for Finances (Conseil d'État et Finances) expedited financial matters of secondary importance, while the Financial Arbitration Court (Grande Direction des Finances) was an administrative tribunal settling disputes between the state and individuals or corporations. Each of these subdivisions of the king's council contained more members than the exclusive High Council, made up of the secretaries of state and of financial and judicial experts.

The initial group composing the High Council contributed a great deal to the basic pattern of Louis's reign, particularly in military, fiscal, naval, and commercial attitudes, partly because many of those who followed as ministers came from the same tightly knit group of royal servants. The five members of the Colbert family have been mentioned; there were also three Le Telliers; and, while only one member of the Phélypeaux family, Louis II, Count de Pontchartrain, was a minister, four served as important secretaries of state. All these counselors reflected the attitude of the king himself: they worked extremely hard; they proffered advice but were under no illusions about the danger of arguing once Louis had made up his mind; and they favoured a protectionist, paternalist policy, whether in the organization of industry, the administration of the colonies, or the building up of the navy. Only toward the end of the reign, with the establishment of the Council of Commerce in 1700, did a less regulatory policy show signs of emerging.

The intendants

To carry out the decisions reached in his intimate and secret High Council, Louis relied chiefly on his provincial intendants. Their powers were completely undifferentiated, and their commissions varied according to changes in royal policy. Like the ministers at the centre, they depended upon the king for their security of tenure. In the provinces they could exercise powers of police; raise military forces; regulate industrial, commercial, and agricultural matters; enforce censorship; administer the financial affairs of various communities; assign and collect taxes; and wield considerable judicial authority in civil and criminal affairs. Inevitably, these agents of the central government created considerable friction and hostility. These new men, with no local roots, answerable only to the king and acting almost invariably in an authoritarian context, were deeply resented by older royal officials, by municipal authorities and guilds, and by local *parlements* and estates—all of whom operated through well-established channels and according to traditional local privileges. The use of intendants, who held neither venal nor hereditary office, was one way in which the limiting effect of the sale of office on royal policies could be circumvented. The authoritarian element of Louis XIV's reign is undeniable: he was determined that no institution or social class would escape the supervision of the crown and its ministers. Thus the power of patronage, which had been exercised for generations in provincial noble households, began to lose its political significance as the king's ministers built up their own alternative administrative clienteles.

In particular, because the Frondes had remained a painful memory from his childhood, the king never allowed the great nobles a similar opportunity for revolt. Versailles became a place of surveillance for pensioned noblemen and their families whose only serious occupation was the traditional one of arms, for the pursuit of which Louis provided ample opportunities. The second rebellious group in the Frondes, the members of the Parlement of Paris, were likewise subjected to stringent controls. In 1673 Louis produced regulations stipulating that the court's remonstrances against royal enactments sent to it could in future only be made after the laws concerned had been registered. By this device the king effectively muzzled the magistrates' criticisms of royal policy. It was equally his intention to overcome the delaying tactics of the provincial courts, especially those situated close to vulnerable frontiers.

Louis's religious policy. Louis was also on his guard against religious dissent. Like most of his contemporaries, he believed that toleration had no virtue and that unity in the state was extremely difficult to maintain where two or more churches were tolerated. Consequently, especially after 1678, Louis intensified the persecution of Protestants; churches were destroyed, certain professions were put out of reach of the Huguenots, and Protestant children were taken away from their parents and brought up as Roman

Revocation of the Edict of Nantes

Catholics. The notorious practice of dragonnades, the billeting of soldiers on Protestant families with permission to behave as brutally as they wished, was introduced. Finally, in 1685 the Edict of Nantes was revoked in order that Louis could claim that he had succeeded where Emperor Leopold I had failed—that is, in extirpating Protestantism from his realm.

The revocation of the Edict of Nantes angered Protestant Europe at a time when Louis's European designs were beginning to meet serious resistance. The revocation deprived France of a number of gifted craftsmen, sailors, and soldiers. At least 600 officers, including Marshal Friedrich, Count von Schomberg, and Henri de Massue, Marquis de Ruvigny (later the Earl of Galway), joined William of Orange, the leader of the Grand Alliance against Louis. Research, however, has reversed the earlier view that the decay of French industry at the end of Louis's reign was the direct result of the expulsion of Huguenot mercantile talent.

The same zeal for uniformity made Louis attack the Jansenists. The theological position of the Jansenists is difficult to define; but Louis, who was no theologian, was content with the simple fact that these zealous Catholics had taken up an unorthodox position that threatened the unity of the state. The movement had begun over the perennial issue of grace and free will as it was propounded in the *Augustinus* of Bishop Cornelius Otto Jansen, published in 1640. In 1653 Pope Innocent X condemned five propositions from Jansen's doctrine, but the movement grew in strength with notable adherents, including Jean-François-Paul de Gondi, Cardinal de Retz, and the great mathematician Blaise Pascal. In 1705 Pope Clement XI published the bull *Vineam Domini,* which further condemned the writings of Jansen; but the archbishop of Paris, Louis-Antoine Cardinal de Noailles, appeared ready to lead the Jansenist forces in opposition to the pope. Under the influence of his confessor, Père Michel Le Tellier, Louis decided to ask the pope for another formal condemnation of the creed. Finally, in 1713, the famous bull *Unigenitus* was promulgated, which, far from ending Jansenism, drove it in the following reign into a disruptive alliance with Gallicanism. Louis's real attitude in this situation is not entirely clear: certainly his policy was in keeping with his authoritarian insistence upon unity. He was suspicious of religious innovation, and his action was consistent with the increasingly orthodox and rigid mood of his last years. Yet, in seeking the pope's support in this matter, he was reversing years of bitter hostility toward Rome when, like many of his predecessors, including Francis I and Henry IV, he had leaned heavily upon the traditional Gallican doctrine.

According to that doctrine, the French king possessed the right of temporal and spiritual regale—that is, the right to nominate new bishops and to administer and draw the revenue from bishoprics while they remained vacant. In 1673 Louis extended this right to the whole of the French kingdom, which had been enlarged in the recent War of Devolution (see below), despite papal opposition. Eventually, in 1682, the Gallican Articles were published as a law of the French state, asserting that the king was in no way subject to the pope in temporal matters and could not be excommunicated and reaffirming the independence of the French church from Rome. The mutual animosity of king and pope only ended in 1693, when, following William of Orange's successful attempt to secure the English throne, Louis agreed to suspend the edict of 1682; but it was a suspension only, not a recantation. The tradition of Gallican independence remained.

Absolutism of Louis. Thus, in religious matters (except where Jansenism was concerned), in his dealings with the nobility and the Parlement, in his attitude to the economy, and in his manner of governing the country, Louis revealed a desire to exercise a paternal control of affairs that might suggest a modern dictator rather than a 17th-century king. Though such a comparison has been made, it is most misleading; neither in theoretical nor in practical terms could Louis XIV be thought of as all-powerful. First of all, the legitimacy of his position under the law—the ancient fundamental law of succession—made him the interpreter of the law and the fount of justice in the state, not a capricious autocrat. Similarly, his kingship bestowed upon him a quasi-spiritual role, symbolized by his consecration with holy oil at his coronation, which obliged him to govern justly in accordance with the laws of God and Christian morality. He was also bound by the need to take counsel; and though he always made up his own mind, he insisted on receiving advice on all important matters of state, which further restricted any arbitrary instincts. Next, there was the essentially federal nature of the country with its collection of such peripheral provinces as Brittany, Normandy, and Provence, all retaining their own Estates and customs. Within both these *pays d'état* and *pays d'élection* (where the Estates no longer met) there was a variety of groups and corporations, not to mention individuals, with their own legally held rights, privileges, and exemptions, such as the nobility, the clergy, the towns, and the king's officers. To impose rigid uniformity in such a situation was both impossible and undreamed of by contemporaries. On the contrary, one of the king's prime obligations was to uphold and respect the myriad different rights to which his subjects laid claim.

Limitation of Louis's power

Louis's financial restraints

Perhaps most of all the king was limited by financial stringency. Louis could and often did try to persuade the cities and provincial Estates to raise their contributions and the clergy to increase the size of their *don gratuit* ("free gift"); he also created more offices and annuities. But these were mere palliatives, and the king was forced on two occasions to introduce novel measures: in 1695 he levied a capitation, or head tax, applicable to all French laymen, even to the princes of the blood; and in 1710 a *dixième* (tenth) that similarly went against the interests of the privileged classes, including the clergy, by requiring a tenth to be paid to the state from all incomes. Significantly, however, Louis made it perfectly clear on both occasions that he recognized the extraordinary and temporary nature of these impositions, made necessary by the pressures of war. It was impossible to be a despot while financial resources were so precarious. The notion persisted that the king should ordinarily live off his own domain. It was also impossible while no nationwide police force existed and while the state of communications remained so poor. All these factors make it clear that a situation simply did not exist in which totalitarian government, at least by 20th-century standards, could have had any meaning.

Finally, Louis XIV remained the prisoner of France's social structure. It is sometimes alleged that the king ruled through the bourgeoisie. It is true that a number of the most distinguished families of the reign were not of ancient nobility, but their faithful and effective service to the king was rewarded in an entirely traditional way—by social elevation. Colbert's father was an unsuccessful merchant, but all his granddaughters married dukes. In other words, the opportunity to enter the highest ranks of the nobility, which had long been available in France, was simply emphasized by Louis XIV. As the greatest nobleman in France, he had no doubt that he must retain the prestige and privileges of the nobility; but he knew equally well that the nobility should not become a caste closed to ambitious and able men. He thus maintained the tradition of royal patronage, which helped to defuse social conflict.

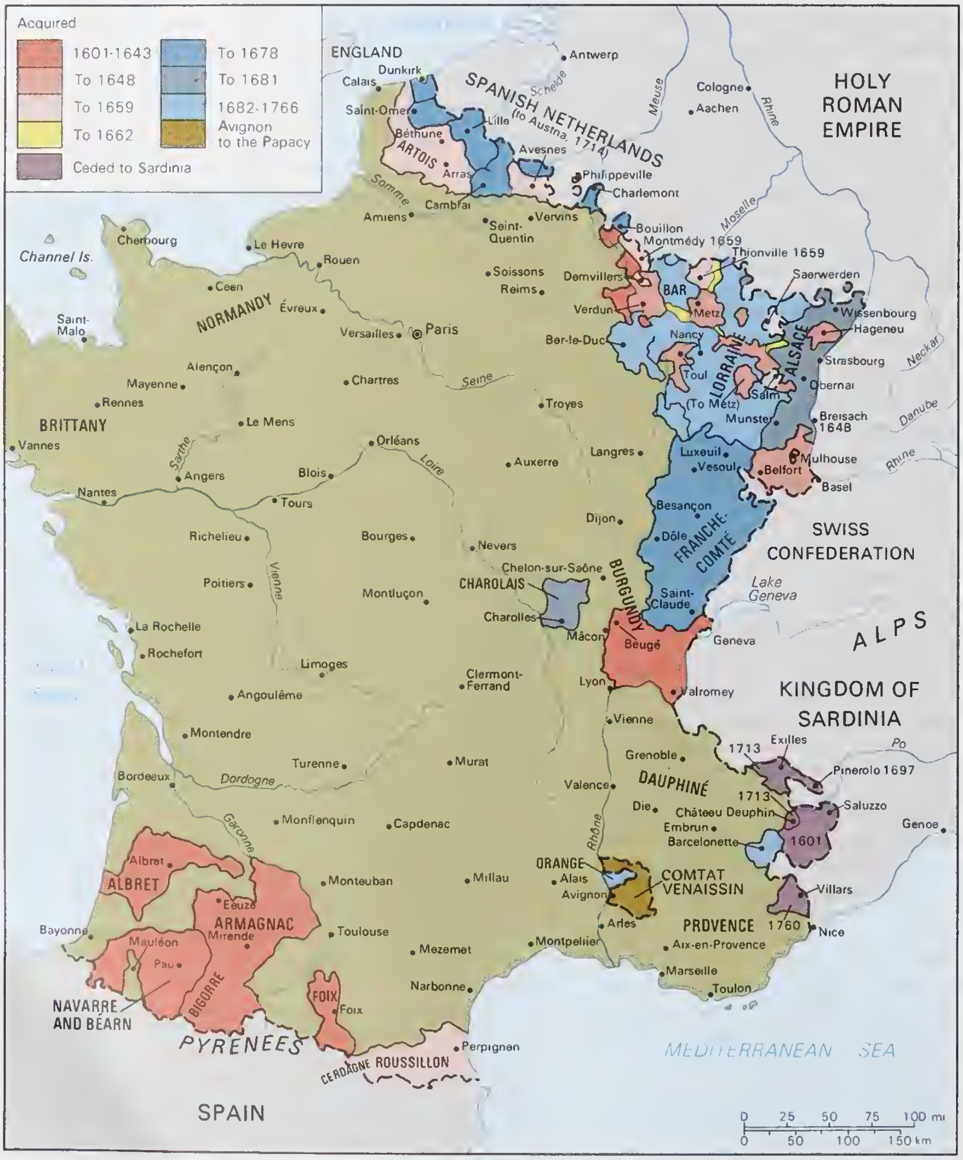
Foreign affairs. From the beginning of his reign Louis pursued a vigorous foreign policy. Historical opinion has traditionally held that Louis sought to dominate Europe, only to meet his just deserts at the end of his reign. More recently another interpretation has emerged, which argues that Louis pursued consistent and for the most part moderate aims and pursued them successfully up to and including the Treaty of Utrecht (1713). (For the traditional interpretation, see the article GERMANY: *The age of Louis XIV.*)

The starting point for the more recent interpretation is the ambiguous Peace of Münster (1648), forming part of the great European settlement of Westphalia, the terms of which subsequently became a bone of contention between Bourbon and Habsburg rulers. One of the critical issues of the treaty was the fate of the three bishoprics of Metz, Toul, and Verdun on the northeast frontier of France. These bishoprics, occupied by the French since 1552, were formally acquired in 1648 together with a number of towns in nearby Alsace. One of the main Habsburg aims in the War of the League of Augsburg, or War of the Grand Alliance (1689–97), and in the War of the Spanish Succession (1701–14) was the restoration of the three bishoprics and the province of Franche-Comté, also on the eastern frontier of France, connecting Burgundy with Alsace, which Louis had acquired at the Treaties of Nijmegen (1678–79) at the end of the Dutch War (1672–78). Louis, however, was determined to hold onto the gains in Alsace, however ambiguously acquired; he also hoped to add Lorraine, to the north of Franche-Comté, to further consolidate this least secure French frontier area. Lorraine was periodically occupied by French troops, notably between 1633 and 1648 and between 1670 and 1697, and Louis sought through various exchange schemes to incorporate the territory into France. The French share of the Partition Treaties (1698, 1700) in Italy was intended as an exchange for Lorraine and not as a means of threatening English and Dutch trade with the Levant. Lorraine, however, did not become a part of France until 1766.

Louis's policy in the northeast was constant and understandable. Franche-Comté was one entry into France previously exploited by its enemies that Louis succeeded in closing in 1678. He had already closed another, the port of Dunkirk, by purchasing it from Charles II of England in 1662; a third gateway, from the southern Netherlands, was effectively barred by the military fortifications erected by his great military engineer, Sébastien Le Prestre de Vauban, in the 1680s. The capture of Lorraine would have bolted yet one more dangerous entry. Of course the situation looked quite different from the Habsburg point of view, especially after Louis's seizure of the key city of Strassburg (French: Strasbourg) in 1681, an episode that goes to the heart of the controversial matter of his reunion policy. Following the successful Peace of Nijmegen, Louis began to employ his own judicial courts to claim sovereignty over all the dependencies of territories that he already possessed in Alsace, Franche-Comté, Metz, Toul, and Verdun. The maneuver enabled him to consolidate his control, especially over Alsace and Franche-Comté, though the legality of the claims to some of the alleged "dependencies" was extremely dubious. There was no legal justification whatever for Louis's greatest coup in the area—the seizure in September 1681 of the independent city of Strassburg. To Louis this key city, the door through which imperial armies could pass (and three times in the recently concluded war had passed) into Alsace, represented a serious threat, for Strassburg was within easy reach of the Danube valley and Vienna. His fears in this area may best be illustrated by his offer during the War of the League of Augsburg to waive his claim to the Spanish succession on condition that Nijmegen be respected; that Lorraine be absorbed into France (with proper compensations elsewhere); and that the Spanish and Austrian lands should not be united under one ruler. The Holy Roman

Seizure of Strassburg

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York, revision Copyright © 1964 by Barnes & Noble, Inc.



French expansion, 1600–1766.

emperor Leopold I immediately rejected these proposals. When the final climactic conflict of the reign, the War of the Spanish Succession, was proceeding so badly, Louis offered to relinquish all the gains he had made from the Spanish inheritance; but he desperately hoped to hold on to Metz, Toul, Verdun, Alsace, and Franche-Comté.

The Dutch War

Louis's attitude toward the Dutch was less moderate and more bullying. His invasion of the Spanish Netherlands in 1667 and the ensuing War of Devolution frightened the Dutch into a Triple Alliance with England and Sweden, which led to the Treaty of Aix-la-Chapelle (1668). Then, in the Dutch War that followed shortly afterward (1672–78), Louis intended to warn the Dutch that France was a serious commercial competitor and to force the Dutch to give him a free hand in the Spanish Netherlands when the issue of the Spanish succession came to the fore. He learned from that war that he could never hope to incorporate a large part of the Netherlands into France against Dutch opposition; but he also continued to fear the manner in which the Dutch might try to influence the government of the Spanish Netherlands for their own economic benefit. Here again was an example of mutual hostility and suspicion in which interpretations of motives in Versailles and in The Hague were diametrically opposed. At the Treaty of Rijswijk (1697) the Dutch gained the right to keep a series of Dutch barrier fortresses within the southern Netherlands as a check against French aggression; it was Louis's seizure of these fortresses in 1701 that precipitated the War of the Spanish Succession (1701–14).

That war has usually been depicted as the most significant element in an assessment of Louis's total foreign policy: for some historians, all his relations with the rest of Europe were geared to this great issue; for others, it was the final misjudgment born of overconfidence, provoked by his own ambitious miscalculations, and destined to ruin France. It is certainly true that the approaching end of the direct ruling line in Spain had interested European rulers for many years, and the Bourbon claim to a share in that rich inheritance—deriving from Louis's marriage to Marie-Thérèse, elder daughter of King Philip IV of Spain—was accepted as a key factor in the situation. In 1668 Louis and Emperor Leopold I had gone so far as to sign a partition treaty, more than 30 years before the death of the last Spanish Habsburg, Charles II. No European statesman was surprised, therefore, at Louis's later concern when, after the signature of the Treaty of Rijswijk in 1697, he undertook negotiations with the English king William III out of which two further partition treaties emerged. The crucial moment came when Charles II's last will was published, offering the Spanish crown, in opposition to the second Partition Treaty, to Louis's grandson Philip, Duke d'Anjou (later Philip V). Louis's decision to accept did not in itself provoke war. (There is evidence, however, that both William and the Dutch believed that Charles's will would name the Habsburg candidate, and they certainly had no intention, in the light of what it did contain, of trying to force the emperor to accept a treaty whose provisions he had always resolutely opposed.) Besides, if Louis had snubbed the Spanish offer, it would have been made to Austria, and the spectre of the restoration of Charles V's empire—probably coupled with French losses on the northeastern frontiers—was intolerable. In addition, Louis had recently made peace after the War of the Grand Alliance, the hardest conflict in which he had so far been engaged, and thus had no illusions about the difficulty of overcoming another coalition under William III's leadership. One may conclude that he did not seek war. But he did make decisions that made war likely, including his recognition of the Old Pretender as James III of England, his unexplained decision to protect his grandson's right to the French throne (he was envisaging not a single, united realm of France and Spain but two Bourbon kingdoms, with the senior heir succeeding in France), his occupation of the barrier fortresses, and his seizure of the monopoly of the Spanish-American trade.

Treaty of Utrecht

When peace was signed at Utrecht in 1713, Louis, despite the disasters of the intervening years, succeeded in holding onto the gains in Europe that he had considered vital throughout his reign, including Alsace and Strasbourg. In addition, his grandson remained king of Spain, despite all of the efforts of the Grand Alliance to replace him by their candidate, the Austrian archduke Charles (as Charles III). It is true that in the darkest time of the war, during the years 1708–10, the desperate king was ready to give up these precious gains and was prevented only by the intransigence of his opponents with their impossible demand that he should himself assist in driving his grandson from the throne of Spain. Likewise, a fortuitous change of government in England in 1710, which ushered in the Tory peace ministry, and the elevation of the Austrian archduke to the imperial title as Charles VI in 1711 weakened the unity of purpose of the Grand Alliance and enabled Louis's most effective soldier, Claude-Louis-Hector, Duke de Villars, to stage a military revival. Therefore, the relatively successful conclusion of the war from France's point of view was not entirely of Louis's own fashioning. Had events forced Louis to accept a total surrender, it would have been even more tempting for historians to blame the defeat upon the excessive ambitions of an arrogant man.

It cannot be denied that Louis was arrogant and that his arrogance aroused fear and resentment in his neighbours. Equally, he was intolerant, like most of his contemporaries, and feared by Protestant powers as the leader of a new and vengeful Counter-Reformation, an irony in view of his secret encouragement of the Turks in order to weaken the emperor. Both facets of the great king need to be borne in mind when assessing his overall foreign policy, and they help to counter any tendency to overestimate the defensive nature of his strategy. That defensive element, however, is of significance and has been largely lost sight of, especially in assessments of the reign written in English. Louis frightened Europe with his quest for *la gloire,* by which he meant the favourable verdict of history on his contribution to French security and territorial integrity but which his enemies interpreted more narrowly as a preoccupation with military triumphs and vainglorious display. That contemporary interpretation, still widely accepted nearly three centuries later, does less than justice to Louis's shrewd appreciation of political realities and of France's long-term interests.

FRENCH CULTURE IN THE 17TH CENTURY

If historians are not yet agreed on the political motives of Louis XIV, they all accept, however, the cultural and artistic significance of the epoch over which he and his two 17th-century predecessors reigned. In their different ways—Henry IV's interest lay in town planning, Louis XIII's in music, and Louis XIV's in the theatre and in landscape gardening—they all actively stimulated the emergence of great talents and were aided by such royal ministers as Richelieu and Mazarin, who were considered patrons in their own right.

The rebuilding of Paris

From Henry IV's reign dates the rebuilding of Paris as a tasteful, ordered city, with the extensions to the Louvre and the building of the Pont Neuf and the Place Dauphine and, outside the capital, the renovations and extensions at Fontainebleau and Saint-Germain-en-Laye. Henry succeeded in making Paris what it had never been before—the centre of polite society—and he must therefore take some credit, though he was not personally interested in such matters, for the establishment of the famous salon of Catherine de Vivonne, Marquise de Rambouillet, which flourished from 1617 until 1665. There, men of letters mingled with the great nobility to the mutual advantage of both. The guests at her salon included the statesmen Richelieu and the Great Condé; the epigrammatist the Duke de La Rochefoucauld; the letter writer Marie de Rabutin-Chantal, Marquise de Sévigné; the novelist Madeleine de Scudéry; the poet François de Malherbe; and the dramatist Pierre Corneille.

Richelieu also was a key figure in the artistic and architectural development of Paris during his years in power. He was fortunate to employ the great architect Jacques Le Mercier, who built for him, close to the Louvre, the Palais-Cardinal, later the Palais-Royal; it contained two theatres and a gallery for the cardinal's objets d'art. Under the same patron, Le Mercier also built the church of the Sorbonne, where Richelieu is buried. In the world of painting, the cardinal supported Simon Vouet, who decorated the Palais-Cardinal, and Philippe de Champaigne, whose surviving portraits include famous representations of Richelieu himself. The cardinal's most notable contribution, however, was in the field of letters, with the establishment in 1634 of the Académie Française to regulate and maintain the standards of the French language. One of its first tasks was the production of a standard dictionary, a massive work published in four volumes in 1694. The Académie succeeded over the years in making the pursuit of letters socially acceptable, though still inferior to the pursuit of arms. Finally, Richelieu's great interest in the theatre persuaded him to patronize a number of dramatists, including Jean de Rotrou and Pierre Corneille. Richelieu's patronage of the arts was taken over by his great pupil Mazarin, who collected some 500 paintings. He housed them in the Palais Mazarin (now the National Library), which itself was enlarged for Mazarin by the architect François Mansart. He also commissioned Louis Le Vau to rebuild part of the medieval castle of Vincennes, thus setting him off on his successful career.

Versailles

Louis XIV's patronage centred on Versailles, the great palace that also played such an important part in the political life of 17th-century France. There André Le Nôtre designed the formal gardens, which still attract a multitude of admiring visitors, as they did when they were first completed. There Jules Hardouin-Mansart added the long, familiar garden facade, and Charles Le Brun decorated the Gallery of Mirrors and the adjoining rooms of war and peace with unforgettable magnificence. There the composer Jean-Baptiste Lully devised and directed a number of musical entertainments with such success that Louis granted him noble status and the office of a royal secretary. There, too, the comic genius Molière was encouraged by the king's support; after the dramatist's death Louis was directly responsible for the establishment, in 1680, of the Comédie Française. There, finally, Louis recognized the genius of Jean Racine, whose great tragedies, from *Bérénice* (1670) to *Iphigénie* (1675), earned him membership in the Académie Française and a noble office, that of *trésorier de France,* from the King.

This blossoming of the arts was aided though not inspired by the patronage of kings and ministers. The artistic creations evince a strong element of order and simplicity culminating in the classical grandeur of Racine's plays and the facade of Versailles. Thus they might seem to reflect the growth of political stability and order over which Louis XIV presided. It is, however, dangerous to tie creative achievements in the arts and sciences too closely to their political environment. Moreover, there are significant counterpoints to the theme of classical order. The philosopher René Descartes's doubting, rationalistic approach to the fundamental questions of God's existence and man's relationship to God undermined the rigid adherence to revealed truths propounded by Bishop Bossuet. The Jansenist Blaise Pascal, one of the most versatile geniuses of the century, represented and defended a minority religious movement that Louis XIV believed dangerously subversive. Toward the end of his long reign, Louis encountered the fierce social criticism of Jean de La Bruyère and the skepticism of Pierre Bayle. These discordant elements draw attention to the fact that the absolute state which Versailles was intended to represent concealed tensions that would surface after the king's death. Nonetheless, the splendour of Versailles and the classical simplicity of Racine's tragedies represent a high point in creative human achievement, and it is to his credit that the king chose to be identified with them. (J.H.Sh.)

France, 1715–89

The year 1789 is the great dividing line in the history of modern France. The fall of the Bastille, the French state prison, on July 14, 1789, symbolizes for that nation, as well as for all other nations, the end of the premodern era characterized by an organicist and religiously sanctioned traditionalism. With the French Revolution began the institutionalization of secularized individualism in both social life and politics; individualism and rationality found expression in parliamentary government and written constitutionalism. Obviously, the English and American revolutions of 1688 and 1776 prefigure these changes, but it was the more universalist French Revolution that placed individualism and rationality squarely at the centre of human concerns.

Because the revolutionary events had such earthshaking power, the history of France in the century preceding 1789 has until recently been seen as a long and quickening introduction, as a period marked by the decay of the *ancien régime* ("old regime"), a locution that came into its own during the Revolution. Some contemporary historians, however, reject this view and present 18th-century France as a society undergoing rapid but manageable change, especially in cultural concerns. They perceive the French Revolution as a political event that could have been avoided if the French monarchy had been more consistent in its effort to modify political institutions in order to keep up with social and especially cultural change.

THE SOCIAL AND POLITICAL HERITAGE

The social order of the ancien régime. To understand the developments of the 18th century and to follow the scholarly debates, one may begin with a definition of the ancien régime. Its essence lay in the interweaving of the state's social, political, and economic forms; the term itself, though primarily a political concept, has also always had a clear and social and economic resonance.

Nature of the ancien régime

In the society of the ancien régime all men and women were by birth subjects of the king of France as well as members of an estate and province. In theory always and in practice often, the lives of French men and women of all ranks and estates took shape within a number of overlapping institutions, each with rules that entitled its members to enjoy particular privileges (a term derived from the Latin words for "private law"). Rights and status flowed as a rule from the corps to the individual rather than from individuals to the group, as was true after 1789.

France itself can be conceived of as an aggregate of differentiated groups or communities (villages, parishes, or guilds), all of them theoretically comparable, but all of them different. In many respects the kingdom was an assembly of varying provinces, a number of them endowed with vestigial representative institutions. In some important ways France was not truly a unit of government. Unlike England, for example, France was not a single customs union; more tariffs had to be paid by shippers on brandy floated down the Garonne to Bordeaux than on wine shipped from France to Britain.

The concept of national citizenship was not unknown in France under the ancien régime, existing in the sense that all Frenchmen, regardless of their rank and privileges, had certain legal rights denied to all foreigners. There was, however, no French nation whose citizens taken one by one were equal before man-made law, as was true after 1789. Laws were in the main inherited, not made. This is not to say that France, though structured around the "premodern" concept of the guild, or group, or *corps,* was a static or, materially speaking, a stable society. For many artisans, peregrination was a way of life, and many years of their young manhood were spent on a *tour de France,* which took them from city to city in order to learn their trade. Serfdom was practically unknown (only 140,000 serfs remained in France in 1789, and none of them on crown lands, where Jacques Necker, the comptroller general, had abolished serfdom in 1779), and peasants were free to move as they wished from one village to the next. Indeed, such large numbers of people were moving around that the fear of unattached vagrants was strong in prerevolutionary France.

Monarchy and church. In the 18th century the king still thought of himself as the feudal suzerain of his subjects. Familial imagery was an important component of royal rhetoric; the king of France was father of his subjects. His right to reign echoed all husbands' right to rule over their wives and all fathers' right to rule over their children. His messages, however draconian and confiscatory they might be, were invariably couched in a rhetoric of religious and paternal solicitude.

King and church

The king, moreover, was a Christian monarch and as such was endowed with quasi-priestly functions. He was anointed at his coronation with holy chrism said to have been brought from heaven by a dove. It was thought that as evidence of his special status he could cure scrofula by his touch. The relationship of church and state was complex. Oftentimes the king did not hesitate to exploit the church, over which he held extensive power by virtue of the still-valid Concordat of Bologna of 1516. Monarchs used their right to appoint bishops and abbots to secure the loyalty of impoverished or ambitious nobles. The king could also extract whatever moneys he wished from the church taken as a whole, and schemes of wholesale confiscation were bruited from time to time. Some monasteries were reformed or consolidated by the crown.

Nonetheless, until 1788 the Roman Catholic church retained in France unusually broad doctrinal rights and social prestige, even by the standards prevailing in central or southern Europe, not to speak of what held true in the far more tolerant countries of northern Europe (Prussia, Holland, and Britain). French Protestants were denied religious toleration until 1787. Jews were only tolerated as quasi-foreigners until 1791. Of considerable symbolic importance was the fact that before 1789 it was the church that kept the registers of births and deaths which marked the beginning and end of each person's earthly existence. The church, the police, and the courts collaborated closely to maintain the prestige of religion; until at least the 1780s the church severely condemned licentious or irreligious books such as Rousseau's *Émile*, which was burned in 1762 by order of the Parlement of Paris (as it also was, incidentally, in Geneva, and in Rome by order of the pope).

The monarchy basically respected the various rights of the church accrued by tradition, as it did the civil and property rights, or "liberties," of its subjects generally. Continuity ordinarily seemed to be the first principle of the French state, and it was inherent in the concept of king itself: the king was held to have two bodies, a physical one, which necessarily decayed, and a spiritual one, which never died. The main purpose of the French state was to defend vested interests—*i.e.*, to maintain continuity rather than to change the existing order.

Commitment to modernization. The great peculiarity of the ancien régime, however, was that traditionalism, though deeply felt, was only one-half of a complicated institutional diptych. When first conceived by Cardinal Richelieu between 1624 and 1642 and developed after him by Mazarin, Colbert, Louvois, and the Sun King, the ancien régime was also committed to a program of modernization. Guided by a modern *raison d'état*, the state was eager to further changes of all kinds. Administratively, its absolutist will, formulated at Versailles in a complex array of governmental councils, was enforced in the provinces by the intendants and their subordinates. The monarchy also protected modern manufacturing and, more desultorily, modern finance. It protected and firmly guided intellectuals through the Académie Française founded in 1635. With greater hesitation the monarchy also staged France's drive to economic and military supremacy not just in Europe but overseas as well, in Canada, India, Africa, and the Caribbean.

Divided in its goals, some of them traditional and others modern, the state was also ideologically double-minded. In the 17th century many intellectuals (some of them clerics like Bishop Jacques Bénigne Bossuet [1627–1704]) developed a Hobbesian justification of absolutist rule, which was renewed throughout the 18th century. Religion and tradition went hand in hand, but absolutist theoreticians went further. They justified the state's right not only to legislate and tax more or less at will but also to imprison arbitrarily without due process of law. The lettres de cachet, which allowed the king to have individuals committed to the Bastille and to other prisons forever and without any kind of trial, were seldom given out, and usually to fathers who wished to correct their wayward children. But they did exist, as liberal or scurrilous propagandists knew full well, sometimes at first hand: about one-fourth of the 5,279 people imprisoned in the Bastille between 1660 and 1790 were connected with the world of the book.

Lettres de cachet

CONTINUITY AND CHANGE

The political history of 18th-century France can be conceptualized in terms of the double heritage and the problems it entailed. The discussion may be linked to two issues: first, the economic transformation of a traditional and essentially agricultural society by both commerce and ideas; and, second, the state's efforts (and eventual inability) to modernize and unify its structure and purpose in order to encompass the changed economic and cultural expectations of the nation's elites.

Economic change in 18th-century France was not uniform. Some of it was of a familiar and recurrent kind, such as increases in population and food shortages, and new and unfamiliar kinds of transformations created by economic and cultural developments. Whereas the first kind of change primarily affected rural France, the second kind was more pronounced in urban France, especially in the north and west, where economic and cultural developments had created a population that longed, often unconsciously, for the deep-seated political upheaval that eventually occurred in 1789.

Comparison with England

Agricultural patterns. In its basic organization, French agriculture continued in its age-old patterns. This contrasted starkly with England, where new agricultural techniques as well as major changes in the control of land—convertible husbandry (a progressive form of land use that did away with the wasteful fallowing of land every two or three years) and the enclosure movement (which made possible the consolidation of small parcels of land into large farms fenced off from use by the rest of the community)—had caused an agricultural revolution. In France there was no significant enclosure movement, despite enabling legislation that allowed the division of some common lands in 1767 and again in 1773. But ordinarily held plots were not divided, and communal patterns of planting—very common in northern France, where a three-field system ordinarily prevailed—were not suspended. Though leaseholding tended to become more frequent, most peasants were sharecroppers, especially in southern France (where a two-field system prevailed). Capitalism did not become a basic trait of rural life; peasants neither produced for the market nor diversified their crops. Their primary goal was not to become richer by taking advantage of economic opportunity but to keep their family afloat by growing enough grain.

Peasants. The condition of many peasants deteriorated markedly in the 18th century; perhaps as many as one-third of them were sporadically indigent. This cannot be explained by a decline in the peasants' share of the land. In 1789, French peasants still owned about one-third of arable land, most of it in small plots of less than 10 acres (nobles owned about one-fifth of the land, the church one-sixth, and bourgeois landlords about one-third as well). In the first half of the century, overall yields may even have risen imperceptibly, and in some years between 1780 and the Revolution perhaps by as much as 1 percent. But the peasant population itself grew more quickly than did agricultural production; thus the per capita production did not change very much and may even have declined. Most contemporaries, including Voltaire, underestimated the population expansion, but it is now known to have been considerable. Fears of scarcity and even starvation were ever present.

Population growth

Demographic changes. Rates of demographic growth varied greatly from place to place: in Brittany, where typhoid fever was endemic, population figures peaked in 1700. In Quercy they grew by 70 percent between 1700 and 1786. But, regardless of local variations, France as a whole counted two million more peasants in 1790 than during the reign of Louis XIV. Better preventive medicine, a decline in infant mortality, and the near disappearance of widespread famine after 1709 all served to increase the size of the population. Birth rates continued to be very high, despite both a traditional pattern of late marriage (men on the average at 27, women at 24 or 25) and the beginnings of the practice of birth control, whose effect was to become evident only after the Revolution. The yearly number of deaths per 10,000 fell from about 400 in 1750 to 350 in 1775, 328 in 1790, and 298 in 1800. In consequence, the average life expectancy of peasants rose from 21 to 27 years (although many individuals, of course, lived longer). Because emigration either to the cities or abroad was slight, the number of rural wage earners grew and did so at a time when prices rose, which they did by more than 60 percent between 1726 and 1789. Material hardship was certainly in 1789 a critical factor behind the deep malaise of the peasantry. Politically, this rural sense of unease mattered a great deal because it made it impossible for the monarchy to rely on the countryside in order to crush urban dissent, as was to happen in central Europe during the Revolutions of 1848.

Rural travails were nothing new; population growth had also been a grievous problem in the 15th and 17th centuries. Still, while significant in themselves, they were all the more troublesome because of other changes concurrently affecting France.

Industrial production. After 1740 overall production in France rose annually by about 2 percent, and even more in some sectors. During the later decades of the 18th century, French industrial production grew rapidly, although not on the same scale as in Britain, whose industrial development had begun 60 years before that of the French. Coal mining was a major industry by 1789, its production being nearly 6 percent higher in the 1780s than in the preceding decade. Mining attracted vast amounts of capital, some of it from the aristocracy. In 1789 the Mines d'Anzin near the Belgian border already employed thousands of workers. In textiles, entrepreneurs like the Swiss Protestant Guillaume-Philippe Oberkampf created new manufactories that permitted better regulation and control of production.

Although France on the eve of the Revolution was not yet a unified national market (as Britain had long since been), largely because internal customs were only abolished after 1789, price discrepancies from province to province, as well as between northern and southern France, were less significant than before. Throughout the country the demand rose for urban manufactured goods and for those luxury items (textiles, porcelains, furniture, *articles de Paris*) in the production of which the French excelled before 1800.

Industrial development

French industrial development was not hindered by the archaic institutional shape of the ancien régime. In 1789 France was only a generation behind England. Although French coal production was not much more than one-tenth that of Britain, the overall rate of French industrial growth in the 18th century was high. Steam-driven pumps, most of them imported from England, were already used in mines and also in Paris, where they served to pump water up from the Seine. French engineers and artisans were highly skilled. French ship design, for example, was superior to that of the English, who routinely copied captured French men-of-war. George Washington, wishing to buy the best watch available anywhere, turned to the American minister in Paris because the world's most accurate timepieces were still made in France.

Commerce. Commerce, especially with the colonies, was an important area of change as well. France's first colonial empire, essentially located in America, was a source of great wealth. Even though France lost both Canada and India during the Seven Years' War (1756–63), the Caribbean sugar islands continued to be the most lucrative source of French colonial activity in the last 100 years of the ancien régime. The French shared the West Indies with Spain and England: Cuba, Puerto Rico, and the western half of Hispaniola belonged to Spain, Jamaica to England; but Guadeloupe, Martinique, and Saint-Domingue (Haiti)—the richest of all nonwhite 18th-century colonies in the world—were French. There 20,000 whites stood an uneasy watch over 160,000 black slaves, imported from Africa, where they had been purchased from black slave merchants who were paid with arms, baubles, and drink. In the islands, the slaves produced sugarcane and coffee, which were refined in France at Nantes, Rochefort, and Bordeaux and often reexported to central and northern Europe. This triangular trade grew 10-fold between 1715 and 1789, and the value of international exports in the 1780s amounted to nearly one-fourth of national income. The sugar trade enriched the planters, the bankers in Paris who had acted as brokers for import and reexport, and the manufacturers of luxury goods that were shipped from France to the Caribbean; the French colonial trade was a closely watched process, governed by mercantilist protective tariffs and rules.

Colonial trade

In the last two decades of the ancien régime the colonial trade lost some of its appeal. Greater profits could then be made in China, India, and East Africa. When the monopoly position of the state-controlled French East India Company ended in 1769, commerce with East Asia was more than doubled. But the West Indies trade remained the single most important aspect of French commercial life in the 18th century. Hence the fact that many of the leading participants of the Revolutionary drama, including the Lameth brothers who dominated the National Assembly in 1791, had connections with the islands. Hence also the appearance of an antislavery movement on the eve of the Revolution. The essayist Charles-Louis de Secondat, Baron de Montesquieu, had already criticized the barbarism of slavery, and in 1788 the Société des Amis des Noirs (Society of the Friends of the Blacks) was founded in Paris and counted among its members not only the Marquis de Lafayette but also Jacques-Pierre Brissot, who later became the leading figure of the Girondin faction, which dominated the Revolutionary government in 1791–92.

Gradually the immoral aspects of trade in the colonies became inescapable, but its first impact was a betterment of economic conditions in France. Obviously, only a few individuals benefited directly from the new currents of exchange, but indirectly millions of Frenchmen were affected by the accelerating tempo of economic life. The circulation of gold specie in the kingdom as a whole rose from 731 million livres in 1715 to some 2 billion livres in 1788. From America to Spain—whose trade with its colonies was under the informal influence and in many instances the informal control of French merchants—silver bullion found its way to northern and eastern France and thence to the eastern Mediterranean, where French influence was widely felt: Bonaparte's Egyptian expedition of 1798 had far-ranging diplomatic, cultural, and economic origins.

Centres of intellectual transformation

Cities. Commerce rather than industry buoyed up French cities, especially the Atlantic seaports. In 1789, 15 percent of Frenchmen lived in cities with more than 2,000 inhabitants. Paris, a city of about 500,000 inhabitants, was only half the size of London, the world's largest seaport. But regardless of their size, French cities were centres of intellectual transformation on the Continent. It was there, in the Sociétés de Pensées, Masonic Lodges, and some 32 provincial academies, that writers found their public. There also took place the cultural revolution that inspired the writers in turn and the economic changes that gave momentum to the cultural upheaval. Although most French cities, especially in central and eastern France, remained unaffected by these changes, many seaports—together with those inland cities benefiting from good avenues of communication—did well or even boomed.

CULTURAL TRANSFORMATION

The Enlightenment. The industrial and commercial developments, already significant by themselves, were the cause, and perhaps also the effect, of a wider and still more momentous change preceding the Revolution—the Enlightenment. Today the Enlightenment can be understood as the conscious formulation of a profound cultural transformation. Epistemologically, the French Enlightenment relied on three sources: rationalism, which had in France a strong tradition dating to Descartes; empiricism, which was borrowed from English thought and which in France underpinned the work of such writers as Claude-Adrien Helvetius (1715–71), Paul-Henri Dietrich, Baron d'Holbach (1723–89), Étienne Bonnot de Condillac (1715–80), and Julien Offroy de La Mettrie (1709–51), the author of a book eloquently entitled *L'Homme machine* (*Man a Machine*); and an amorphous concept of nature that was particularly strong in the immensely popular and important work of Jean-Jacques Rousseau (1712–78) and, in the 1780s, in the works of widely read pre-Romantic writers like Jacques-Henri Bernardin de Saint-Pierre (1737–1814).

Though far apart from one another in a strict philosophical sense, these sources of inspiration generated a number of shared beliefs that were of obvious political consequence. The enlightened subjects of Louis XV and Louis XVI were increasingly convinced that French institutions of government and justice could be radically improved. Tradition seemed to them an increasingly inadequate principle to follow in such matters. Meliorism, gauged especially by the progress of the sciences, was one of the cardinal beliefs of the age. Regarding the economy, physiocrats like the king's own doctor, François Quesnay (1694–1774), praised the virtue of free-market economics and, as they put it, of "laissez-faire, laissez-aller." Widely associated with the humane and rational criticism of religious intolerance and of judicial error were the names of Montesquieu, of the encyclopaedist Denis Diderot (1713–84)—from the first French *Encyclopédie,* published from 1751 to 1772—and, most importantly, of Voltaire (1694–1778). Especially renowned was Voltaire's defense of the Protestant merchant Jean Calas, unjustly broken on the wheel in 1762 for the supposed murder of his suicidal son.

The influence of Rousseau. Regarding the cultural transformations of 18th-century France, it would be difficult to overestimate the crystallizing and revolutionary impact of Rousseau's vision of the world. His thought changed conceptions both of man's private and public spheres. According to Rousseau the self becomes empowered in private union with the beloved other, as portrayed in his immensely popular novel *Julie: ou, la nouvelle Héloïse* (1761; *Julie: or, The New Eloise*), or in public union with one's fraternally minded fellow citizens, as explained in *Du contrat social* (1762; *The Social Contract*), a work less well-known but even more symptomatic of change.

The cultural transformation affecting the domestic sphere ennobled the roles of women as wives and mothers. The nature of masculine and feminine sexuality became a favoured topic for writers, doctors, pamphleteers, or pornographers (some of whom, like Restif de la Bretonne, or even Diderot, were in other respects writers of considerable stature). Moralists reflected anew on the responsibility of children to their parents and on the obligation of mothers to breast-feed their children. At this time the majority of babies born in the cities were sent to be wet-nursed in the country; indeed, more than 80 percent of babies born in Paris in the 1780s were sent out to be suckled by professional nurses. The appeal to mothers to nurture their own children came from Rousseau himself.

Education

Pedagogic questions were widely discussed, and the possibility of creating national systems of education acquired sudden importance. The French administrator, reformer, and economist Anne-Robert-Jacques Turgot, Baron de l'Aulne (1712–81), expressed the new sensibility when he wrote that the education of children was the basis of national unity and mores. The whole emphasis of education shifted. Whereas education before had as its primary purpose the adaptation of the young to established patterns of culture and thought, it now placed a greater emphasis on the individual desires and propensities of children.

In 1763 an otherwise reactionary *parlementaire* named La Chalotais even put forward a scheme for lay and national primary education. An important landmark in this respect too was the expulsion from France in 1764 of the Jesuits, who had theretofore dominated French secondary education. Increasingly, the French language was substituted for Latin in the secondary schools, or *collèges* (the forerunners of today's lycées). Rhetoric gave way to an emphasis on more "natural" manners and modes of expression. History was raised to the level of a serious discipline; with Voltaire's *Siècle de Louis XIV* (1751; *Age of Louis XIV*), modern French historiography began, and there were echoes of this new attitude in the programs of the secondary schools, which added mathematics, physics, and geography to their curriculum.

The 18th-century cultural transformation also redefined the public roles of men as would-be citizens of a harmonious polity. In 17th-century Jansenist thought, man was seen as a guilt-ridden creature who desperately needed God's grace in order to realize his true and hidden self. This vision of man as fallen and of God as hidden had drawn the French bourgeoisie toward introspection and away from politics. It had made absolutism possible. But, in the prerevolutionary decades of the 18th century, Rousseau, whose writings best expressed the transformed sensibility, transcended this cultural inheritance. In his understanding, man was still a fallen creature because society, through which he was fated to realize his nobler self, had drifted away from primitive simplicity. Man, however, might potentially be good by nature. Thus, though presently fallen, he might save himself by simplifying his private life and above all by purifying the public, social, and political institutions affecting him. In Rousseau's imagination, self-realization was to take place within the world and not against it, as Jansenists like Pascal had proposed.

Self-realization within the world

Exposure to writers like Diderot, Guillaume-Thomas-François Raynal (1713–96), author of an anticolonialist *Histoire des deux Indes* (1770; *History of the Two Indies*), and Jean-Jacques Barthélémy (1716–95), to painters like Jacques-Louis David (1748–1825) and Joseph-Marie Vien (1716–1809), to musicians like Christoph Gluck (1714–87), and to visionary architects like Claude-Nicolas Ledoux (1736–1806) and Étienne-Louis Boullée (1728–99) enabled the educated public of the 1770s and '80s to pursue and sharpen their new insights. It allowed them to explore the limits of the private domain as well as to clarify their new understanding of the public good. These radical ideas had transforming power. Rousseau's message especially appealed to the deeper instincts of his contemporaries, inspiring them with a quasi-utopian view of what might be done in this world.

The ideological or cultural transformation was in some ways limited to a narrow segment of society. In 1789 only one-third of the population, living for the most part in northern and eastern France, could both read and write French. (Outside the aristocracy and upper bourgeoisie, literacy for women was considerably below that of men.) About one-third of the king's subjects could not even speak French. Nonetheless, even though probably not much more than half a million people were directly involved in the cultural upheaval, their influence was decisive.

Relaxation of censorship

The concerns of the new "high culture" were intensely personal and, for that reason, deeply felt, even by people who did not participate in it directly. Readers of sentimental prose might after all also be employers, husbands, and fathers, who would treat their dependents differently. In addition, the means through which new cultural forms were diffused were often original, and in any case, sustained. Between 1723 and 1789 more than 30,000 books were published. Newspapers, some of them from abroad, were widely read (and manipulated by the royal government in order to influence opinion). Many pamphleteers were ready to be hired by whoever had money to pay for their services. Lawyers published their briefs. Theatrical performances like Pierre-Augustin Caron de Beaumarchais's comedy *Le Mariage de Figaro* (1784; *The Marriage of Figaro*), which openly exposed aristocratic privilege, were widely publicized events. In the 1780s censorship became increasingly desultory. Public opinion, whose verdict was identified by the middle class not with the expression of its own particular desires but as the voice of universal common sense and reason, became a tribunal of ideological appeal, an intellectual court of last resort, to which even the monarchy instinctively appealed.

These sweeping changes had created a country that by 1788 was deeply divided ideologically and economically. The salons of Paris, many of them directed by women, were the worldwide focus of a rationalist and atheistic Enlightenment; both Catherine the Great and Thomas Jefferson, though far removed from each other in most respects, shared an abiding interest in the latest intellectual fashions from Paris. But, whatever held true for influential circles, most Frenchmen in these same years remained deeply religious, certainly in the provinces but possibly in Paris as well. Most of the books and pictures Parisians bought on the eve of the Revolution were still related to religious themes. The country was also divided economically; whereas France's foreign trade was very lively, most of the rural communities were, by English standards, unproductive and immobile villages.

THE POLITICAL RESPONSE

The historical debate. In broad terms, 18th-century French politics could be defined as the response of the monarchic state to the emergence of the new cultural and economic configurations that had transformed the lives and especially the imaginations of French men and women. The question was whether the Bourbon monarch could rationalize its administration and find a way to adapt itself in the 1770s and '80s to the new perception of the relationship between citizen and state as it had come to be defined by the Enlightenment, especially by Rousseau.

On the issue of political mutation historical opinion is divided. One set of discussions revolves around the issue of whether the monarchy's efforts at modernization were sufficient; whereas some historians believe that the ancien régime almost succeeded, first in the 1770s and once again in the early 1780s, others argue more pessimistically that the efforts of the monarchy were insubstantial. A more radical view, by contrast, holds that the extent of reform was irrelevant because no monarch, however brilliant, could have met the rising liberal and nationalist expectations of tens of thousands of dissatisfied and vocal people, steeped in Enlightenment thought, who were committed to becoming the empowered citizens of a fraternal state.

The weight of evidence appears to be that the monarchy was by the late 1780s doomed to destruction, both from its inability to carry on the absolutist, administrative work formerly accomplished by men like Colbert and by the nature of its critics' desires; the gap separating the Roman Catholic traditionalism of the monarchy and the neoclassical ambitions of nascent public opinion was too wide.

Assessment of foreign policy

Foreign policy and financial crisis. The monarchy's international performance from 1715 to 1783 was basically creditable. From the time of the War of the Spanish Succession (1701–14), when France had been invaded and nearly beaten, French statesmen pursued a double goal—the preservation of the balance of power in Europe, and, in the world at large, the expansion of the French colonial empire and the containment of England. To achieve the first goal, France, during the War of the Austrian Succession (1740–48), sided with Prussia and against Austria, which was then thought the more dangerous power. Rather reluctantly, during the Seven Years' War, France reversed its position and fought against Prussia, whose efforts at aggrandizement had become a threat to the status quo. In the 1770s and '80s French diplomacy worked to preserve both the continental status quo and the continued independence of the smaller European states, many of which were its clients or allies—notably Poland, Sweden, the United Provinces, and Turkey. In this the French were reasonably successful; although they suffered reverses in Poland, which was partitioned, and in the United Provinces, where the Anglo-Prussian Conservative Party with Prussian military help overcame the pro-French liberal groups in 1787, the balance of power in Europe was preserved.

French successes vis-à-vis England, however, were checkered. Efforts to extend French rule in India or to preserve it in Canada were more or less successful in the 1740s but, as stated above, were totally unsuccessful during the Seven Years' War. During that conflict France suffered grievously in spite of the entente between France and Spain, both ruled by Bourbon monarchs, and in spite of the Austrian alliance. France was unable to prevail either on the Continent or in the colonies and, for having tried too much, it failed everywhere. French influence was at a very low ebb in 1763. During the U.S. War of Independence (1775–83), however, the French were most careful not to be also involved in the Austro-Prussian "Potato War" (War of the Bavarian Succession) of 1778–79, and, by concentrating their efforts for the first time on maritime matters, they succeeded in holding and even reversing the advance of England. A contemplated invasion of England did not take place, but for a short while the French navy had control of the high seas. The real victor of the battle of Yorktown, Pa. (1781), in which the British were defeated, was less General George Washington than Admiral François-Joseph-Paul de Grasse (1722–88), whose fleet had entered Chesapeake Bay. Thus, on balance, France more or less succeeded in holding its own in Europe while preserving its colonial empire.

The costs of war

Nonetheless, regardless of defeat or victory, colonial and naval wars were problematic because of their prohibitive cost. In Bourbon France (as in Hanoverian England and the Prussia of the Fredericks) a high percentage of the governmental income was earmarked for war. Navies were a particularly costly commodity. The crown's inability to manage the ever-swelling deficit finally forced it to ask the country's elites for help, which, for reasons unrelated to the various wars and conflicts, they were utterly unwilling to extend unconditionally. Money thus was a large factor in the collapse of the monarchy in 1789.

Ultimately, to be sure, it was not the crown's inability to pay for wars that caused its downfall. Rather, the crown's extreme financial difficulties could have led to reforms; the need for funds might have galvanized the energies of the monarchy to carry forward the task of administrative reordering begun during the reigns of Louis XIII and Louis XIV. A more determined king might have availed himself of the problems raised by the deficit in order to overwhelm the defenders of Roman Catholic traditionalism. In so doing, the monarchy might have satisfied enough of the desires of the Enlightenment elite to defuse the tense political situation of the late 1770s and '80s. Although in 1789 a program of "reform from above" was no longer possible, it might well have succeeded in the early 1770s.

Domestic policy and reform efforts. As stated above, in the context of 17th-century absolutism, Louis XIV had already initiated many rationalizing reforms. The crown had worked selectively to replace corporatist, traditionalist institutions with meritocratic associations. In 1648, for example, Cardinal Mazarin had already sponsored a Royal Academy of Painting and Sculpture (because it lent itself better to control by the state) as an alternative to the traditional guilds and corps. This statist and anticorporatist program was now embraced, but in a more liberal register, by the Enlightenment partisans of meritocratic individualism. Though Montesquieu had defended intermediary bodies such as guilds as guarantees of civic liberty, thinkers of the Enlightenment attacked them in the name of public utility and of what would later be called the rights of man. In an article written for the *Encyclopédie,* Turgot denied the sanctity of what he called foundations: "Public utility is the supreme law, and cannot be countervailed by a superstitious respect for what has been called the intents of the founders." Most foundations, he thought, had as their only purpose the satisfaction of frivolous vanity. At the other end of the social spectrum, the Protestant Rabaut Saint-Étienne, later president of the National Assembly, argued that "every time one creates a corporate body with privileges one creates a public enemy because a special interest is nothing else than this." No distinction was made between private interest and factional selfishness; in 1786 the future Girondin leader Jacques Brissot was expressing what had become a commonplace when he wrote that "the history of all intermediary bodies proves, in all evidence, that to bring men and to bind men together is to develop their vices and diminish their virtues." Private benevolence applied to public purpose was loudly praised in the 1780s, and Louis XVI's finance minister, Jacques Necker, did a great deal for his reputation by endowing a hospital for sick children, which stands to this day. By 1789 public and charitable concern had become the themes of countless didactic works of literature and painting.

The monarchy's reform efforts

Many of the monarchy's efforts to institutionalize this new sensibility were often significant. The crown encouraged not only agriculture but also manufacturing and commerce. It allowed tax exemptions for newly cultivated land. It subsidized the slave trade, on which much of the prosperity of the Atlantic seaports was based. It improved communications and in 1747 founded a School of Bridges and Roads to train civil engineers for the royal engineering service that had existed since 1599. In the provinces, in Bordeaux especially, many intendants took an active role in road building and in the modernization of urban space. The crown's administrators also gave sustained thought to the abolition of internal customs and to the creation of what would have been the largest free-trade zone in Europe at the time. Social mobility was made possible; after 1750 many successful merchants and bankers were ennobled.

These were important steps. But the royal bureaucrats tried to go much further in regard to both the rationalization of the state's financial machine and the meritocratic individuation of social and economic forms.

Tax reform. In 1749–51 Jean-Baptiste de Machault d'Arnouville, then comptroller general of finances, proposed a partial reform of the tax system, his particular concern being to restrict the financial power of the church. In 1764 and 1765 another comptroller general, François de L'Averdy, attempted a reform of municipal representation and administration. All royal officials understood the need to reform and rationalize both the imposition and the collection of taxes; many nobles were exempted from taxation, especially in northern France, and many taxes were inefficiently collected by private tax-farmers.

The country's overall fiscal structure was highly irrational, as it had been developed by fits and starts under the goad of immediate need. There were direct taxes, some of which were collected directly by the state: the taille (a personal tax), the capitation, and the *vingtième,* which were a form of income tax from which the nobles and officials were usually exempt. There were also indirect taxes that were paid by everyone: the salt tax, or gabelle, which represented nearly one-tenth of royal revenue; the *traites,* or customs duty, internal and external; and the *aides,* or excise taxes, levied on the sale of items as diverse as wine, tobacco, and iron. All the indirect taxes were extremely unpopular and had much to do with the state's inability to rally the rural masses to its side in 1789. In the 1740s attempts had been made to amend this system but had foundered on the *parlements'* opposition to a more equitable distribution of taxation. By 1770 the swelling debt made it obvious that something should be done. Unpopular measures, such as forced loans, were put into effect. Joseph-Marie Terray, Louis XV's comptroller general of finances, repudiated a part of the debt.

Indirect taxes

Some observers, partisans of enlightened despotism (like Voltaire, who defended it indirectly in his play of 1773 entitled *Les Lois de Minos* [*The Laws of Minos*]) understood that the French monarchy stood in this particular instance for administrative rationalization and progress. But the current of opinion was already moving against the crown. Many writers saw in Terray a tool of royal despotism, plain and simple, and his ministerial colleague René-Nicolas de Maupeou (1714–92) was even more detested for his destruction of the *parlements,* an unusual institution that had become the bastion of conservative opposition to royal reform.

Parlements. The 13 *parlements* (that of Paris being by far the most important) were by their origins law courts. Although their apologists claimed in 1732 that the *parlements* had emerged from the ancient *judicium Francorum* of the Frankish tribes, they had in fact been created by the king in the Middle Ages to dispense justice in his name. With the atrophy of the Estates-General, which had not met since 1614, the *parlements* now claimed to represent the estates when those were not in session. In 1752 a Jansenist *parlementaire,* the Abbé André-René Le Paige, developed the idea that the various *parlements* should be thought of as the "classes" or parts of a larger and single "Parlement de France."

This was a politically significant claim because these courts had taken on many other quasi-administrative functions that were related to charity, education, the supervision of the police, and even ecclesiastical discipline. Royal decrees were not binding, claimed the *parlementaires,* unless the *parlements* had registered them as laws. Although the *parlementaires* admitted that the king might force them to register his decrees by staging a *lit-de-justice,* (*i.e.,* by appearing in person at their session), they also knew that the public deplored such maneuvers, which manifestly went against the grain of the monarch's supposed Christian and paternalist solicitude for the well-being of his subjects.

Various social, cultural, and institutional developments had served to turn the parlements into strongholds of conservative resistance. Since the 17th century the monarchy's need for money and the ensuing venality of offices had enabled the *parlementaires* to buy their offices and to become a small and self-conscious elite, a new "nobility of the robe." In 1604 the creation of the *paulette* tax had enabled the *parlementaires* to make their offices a part of their family patrimony, even if the value of their offices fell somewhat during the course of the 18th century. They had gained status by intermarrying with the older chivalric nobility of the sword. By 1700 the *parlementaires* had become a hereditary and rich landowning elite. (Near Bordeaux, for example, the best vineyards were theirs.) The interregnum of the regency after the death of Louis XIV (1715–23) had given them a chance to recapture some of the ground they had lost during the Sun King's reign; the value of their offices, however, fell again somewhat in the course of the 18th century. The *parlementaires's* Jansenist leanings before 1750 and their recent espousal of antiabsolutism—expressed in the work of Montesquieu, himself a baron and a *parlementaire*—gave this elite ideological consistency.

The *parlementaires*

In 1764 the Jansenist *parlementaires,* as ideological "progressives," secured the expulsion of the Jesuits from France. Then in 1766, as defenders of the Christian social order, they also secured the execution of the 18-year-old Chevalier de la Barre, accused of mutilating a crucifix and owning a copy of Voltaire's *Philosophic Dictionary.* In 1768–69 the Parlement of Brittany, in an antiabsolutist stance, forced the resignation of an appointed royal official, the Duke d'Aiguillon, who had dared to try to limit the power of the local nobility, with whom the Parlement was now in close alliance.

King and *parlements.* In 1770 the *parlements* had reached such power that Louis XV was finally goaded into a burst of absolutist energy. The Paris Parlements, which had dared to attack Terray's financial reform, were dissolved on Jan. 19, 1771. Maupeou was then authorized to create an altogether different set of *parlements* with appointed judges shorn of administrative and political power.

In time, opinion might well have accepted Terray's and Maupeou's reforms. France might then, like Prussia, have avoided revolution from below through the practice of a revolution from above. But the death of Louis XV in 1774 put an end to the experiment. His 20-year-old successor, Louis XVI (ruled 1774–92), unsure of himself and eager to please, recalled the *parlements.* To make matters worse, an indirect repetition after 1774 of this same cycle, with the renewed victory of the *parlementaires* over the enlightened royal bureaucracy, set the seal of doom on the absolutist ancien régime.

In late 1774 Louis XVI appointed Turgot, a former intendant, comptroller general. Perhaps because he thought that the success of his reforms would guarantee their acceptance, perhaps also because he thought it vain to attack the Parlement directly so soon after Maupeou's dismissal, Turgot carried through his measures without first destroying the institutional bases of privileged conservatism. He left the Parlement alone and attempted instead to reduce government expenditures and to alter the methods of tax collecting. In accordance with his physiocratic laissez-faire principles, he freed the corn trade from restraint; he suppressed the corvée, or forced labour service, exacted from the peasants; and abolished the guilds, which had limited both access to artisanal professions and the competition within them. Finally, he suggested that Protestants should be given freedom of conscience. In short, Turgot attempted to rationalize the administrative practices of the French state and to individuate French social and economic life. The solution to the financial crisis, he thought, would come not through the state's appropriation of a larger share of extant resources but from the expansion of the nation's ability to produce and pay. The strength of creative individualism, he thought, would break the political impasse.

Turgot's reforms

In May of 1776, however, Turgot was dismissed. Opposition to his measures had come from all sides: a poor harvest had sparked peasant disturbances, the clericalists were antagonized by Turgot's philosophical friends (his greatest and most loyal disciple was Marie-Jean-Antoine-Nicolas de Caritat, Marquis de Condorcet, the future

Girondin), and, when the Parlement of Paris once again refused to register the new edicts, Louis abandoned Turgot as he had dismissed Maupeou. Thenceforth, the state carried through only minor reforms, none of them on a scale commensurate with the needs felt by the Enlightenment bourgeoisie and notables of the cities and towns. The vestiges of serfdom were suppressed in 1779, and restrictions were removed from the Jews in 1784. In 1786 the salaries of the poorer country clergy were raised, and there were also improvements in the organization of the navy and the army (especially of the artillery, in which young Napoleon became a second lieutenant in 1786). Judicial procedure was also revamped: in 1780 torture was abolished, and in 1784 the king's use of lettres de cachet for purposes of arbitrary imprisonment without trial was considerably curtailed. But these were minor adjustments. Nothing was done to solve the fundamental problems of the organization of society and of the state in a manner that would be acceptable to progressive public opinion.

By the mid-1780s the state's position had in fact become truly hopeless. Resentment had risen to the point where no concession would suffice; as a gift of the crown, no reform could be acceptable. Most remarkable was the lack of success of deliberative provincial assemblies, elected by large landowners without regard to social class; their function was to help forge the consensus necessary for reforms, especially in the area of taxation. They would have been seen as an incredible concession 20 years before; in the 1770s and '80s they were failures. The gap between the state and the public was so great that the state could now do nothing right.

But the problem of the state was complicated by the fact that in the 1780s most of what it did was actively and positively wrong. Having alienated conservative opinion by its attempted reforms and enlightened opinion by its failure to push them through, the crown had no constituency on which it could rely; its only strategy was to keep things as they were, in the hope of proceeding somehow from expedient to expedient. The U.S. War of Independence, however, was a most inexpedient event. This was not due to its military denouement, for, as stated, in spite of some naval reverses in 1783, the French army and navy performed creditably during that conflict, and the position and prestige of France in Europe was greater in 1783 than it had been for 40 years. The French problem in the war was not military but financial. The cost of the war was enormous, and it was met not by new taxation but by new borrowing, which doubled the size of the national debt. In the late 1780s the service of the funds absorbed more than half the crown's income: by 1789 debt payments had risen to 300 million livres from less than 100 million before the American war. This could not go on forever, and indeed it was only because of the talent of Jacques Necker (Turgot's successor) in raising loans that the system went on as long as it did. But sooner or later, from sheer financial exhaustion, the state would have to turn to the nation, and at that point the whole set of conflicts in society—between country and city, bourgeois and noble—would have to be resolved.

Cost of the American War

Loss of prestige. The absolutist principle of monarchy seemed exhausted, as was the crown's traditionalist and religious prestige. The prestige of Louis XIV had been immense; when Voltaire wrote of the "grand siècle," it was the 17th century he meant and not his own. Gradually, however, the monarchy had lost its sacred aura.

Genuinely popular in the earlier years of his reign, Louis XV had become thoroughly discredited by the time of his death. Although his liaison with Mme de Pompadour had borne many fruits in the form of the patronage she extended to artists as different as Voltaire and the architect Jacques-Ange Gabriel, it did nothing for the prestige of the crown. More unpopular yet was Louis XV's last mistress, Mme du Barry, a former prostitute. The king was also thought to maintain a private brothel (which was true) and to have literally starved his people by speculating on a rising price for grain (which was false).

It is difficult to know if Louis XV's lack of popularity was due to his own actions or to a decline in the prestige of the monarchy. By contrast, Louis XV's grandson and successor, Louis XVI—well-meaning, obtuse, and sexually incompetent—was initially more popular; this would seem to suggest that the crown in the 1770s could still tap the traditional sources of loyalty. But by the 1780s Louis XVI was held in contempt as well; his wife, Marie-Antoinette, was also much hated. Never popular because she was by birth a member of the Austrian ruling house, which in France was traditionally disliked, she reached in 1785 a nadir of unpopularity during the prosecution of the Affair of the Diamond Necklace. (This case revolved around a cardinal, the Prince de Rohan, who had been tricked into giving a diamond necklace to a woman whom he thought to be the queen; in this instance Marie-Antoinette was blameless, but what was seized upon by public opinion was that a cardinal had thought it possible to seduce and bribe the queen.) The discredit that befell the monarchy in consequence was immense, and Napoleon dated the beginning of the French Revolution from this very episode.

Loss of control. By August 1786 even Necker's successor, the frivolous if intelligent Charles-Alexandre de Calonne (1734–1802), had concluded that the French state was coming to a dead end. Although the crown had assured the moneymen that it would not default on the debt, as indeed it could not have done without alienating thousands of its natural supporters, Calonne found that he could no longer raise funds by borrowing. To resolve the ensuing impasse, in February of 1787 he summoned to Versailles the Assembly of Notables, which represented the privileged orders—great nobles, bishops, *parlementaires*—in the hope that they would finally and of their own accord undertake the long-resisted transformation of the state. But the notables refused again, as the *parlementaires* had in 1776, this time on the grounds that Calonne's mismanagement was the only cause of the crown's problems. They argued that if Calonne were replaced by Necker, who had borrowed so successfully, all would be well. Once again, the crown yielded. Calonne was replaced by Loménie de Brienne, a cardinal, though an atheist, and a spokesman of the privileged majority of the Assembly of Notables. Almost at once Loménie reversed himself and came to Calonne's conclusion: the state could not go on as it had. The notables, however, refused to be more amenable to Loménie than they had been to Calonne. Despairing of securing the consent of the privileged orders, Loménie dismissed the assembly in May of 1787, and in August the Paris Parlement was exiled to Troyes.

Opposition of the notables

But these measures were desperate, and already the monarchy was beginning to lose control of the political process. Indeed, for the next two years it merely floundered from one scheme to another, in the impossible hope of squaring the circle of modernistic reform, popular hostility, respect of privilege, and the preservation of royal absolutism. Essentially unwilling to force the privileged notables to yield their corporate rights, the crown was unable to assert any coherent policy. The Parlement was therefore recalled from Troyes in September of 1787, again dismissed in May of 1788, and, in the face of a beginning of a breakdown of law and order and of the inability of officials to collect taxes, once more recalled to Paris by the crown in May of 1788.

Paradoxically, the return of the reactionary *parlementaires* was saluted as a great victory by their most bitter reformist opponents. This is to be explained by the fact that the crown was now perceived by all Frenchmen as inimical: the conservative *parlementaires* opposed it from their reactionary point of view, and the reformists opposed it for its arbitrary repression even of the reactionaries. The crown had become essentially powerless, and on Aug. 8, 1788, it tacitly accepted this fact by agreeing to call together the Estates-General on May 1 of the following year. At that juncture the monarchy in fact dropped out of the political debate, which would now be structured by the tensions within French society. Censorship was in practice suspended (formally in mid-May 1789), and thousands of pamphlets circulated in Paris.

Composition of the forthcoming Estates

The popularity of the Parlement was, however, shortlived. Opinions were polarized regarding the composition of the forthcoming Estates and the procedures to be followed. Debate revolved around the issue of whether they should meet in three groups of equal size (a reinscription of traditional modes of representation) or whether the Third Estate should be doubled in size and the deputies meet together (a step toward the recognition that sovereignty lay with the representatives of the people and not the king). The Parlement of Paris, which had returned in triumph on Sept. 23, 1788, decided on September 25 that the three Estates should meet as before. At once the popularity of the *parlementaires* collapsed. On September 23, they had been seen as victims of royal oppression. Two days later, they were perceived to be what they had in fact long since become, the defenders of the traditional social and political fabric of the ancien régime. In a last and fitful assertion of authority, at the behest of Necker, once again minister, the crown decided on December 27 to overrule the Parlement. The Estates, it resolved, would meet separately, but the Third Estate would have as many deputies as the other two orders combined. The stage was set for the coming Revolution.

THE CAUSES OF THE FRENCH REVOLUTION

In an immediate sense, what brought down the ancien régime was its own inability to change or, more simply, to pay its way. The deeper causes for its collapse are more difficult to establish. One school of interpretation maintains that French society under the ancien régime was rent by class war. This position implies that the French Revolution revolved around issues of class; it has led to the class analysis of prerevolutionary society as well as to the class analysis of the opposing Revolutionary factions of Girondins and Montagnards and, more generally, to what the historian Alfred Cobban called "the social interpretation of the French Revolution."

In keeping with this interpretation, Marxist historians of the 1930s emphasized that the French 18th-century bourgeoisie had assumed a distinct position in French society in that it was in control of commerce, banking, and industry. Revisionist historians in the 1980s, however, responded that the bourgeoisie had no monopoly in these sectors; nobles were also heavily involved in foreign trade, in banking, and in some of the most modern industries, such as coal mining and chemicals.

Most historians today would argue that on balance it was becoming increasingly difficult to distinguish clearly between the nobility and the bourgeoisie. Like most nobles, wealthy French nonnobles were landlords and even owners of seigneuries, which were bought and sold before 1789 like any other commodity. Although one can speak of a secularized "bourgeois" ethic of thrift and prudence that had come into its own (symbolized by the oath, the opposite of the aristocratic and ephemeral fireworks), supporters of this ethic, as of the Enlightenment ethic, were both noble and nonnoble.

Aristocratic privileges

There were two areas, however, in which the nobility enjoyed important institutional privileges: the upper ranks of the army and the clergy were, in the main, aristocratic preserves and had become more so in the 1780s. Henri de Boulainvilliers, in his posthumous essays of 1732 on the nobility of France, had even developed a wholly fraudulent but widely praised theory of noble racial superiority. Thus, there were some issues on which all of the bourgeoisie might unite against most of the nobility. But such issues, it is now claimed, were relatively unimportant.

Contemporary historical work has refocused the discussion regarding the causes for the Revolution. Studying the representation of politics, the shape of Revolutionary festivals, the Revolutionary cults of sacrifice and heroism, scholars have come to place the transformation of culture at the core of their discussion. What really mattered was the desacralization of the monarchy, the new understanding of the self and the public good, and the belief that thinking individuals might seize the state and fundamentally reshape it. Other historians, by contrast, have emphasized the persistent liabilities that French political culture carried through the Enlightenment, such as the suspicion of dissent and the readiness to rely on force in order to subvert it.

From either of these two perspectives, it follows that the prospects of the monarchy's survival were dim in 1788. Many government officials, it is true, were finely attuned to public opinion. The vast neorepublican canvases of Jacques-Louis David, such as his "Oath of the Horatii" (1784) glorifying traditional republicanism, were commissioned by the king's dispenser of patronage, the Marquis d'Angivillers, a friend of Turgot. Visionary architects, developing a style of Revolutionary Neoclassicism, similarly received royal commissions for new public works. Chrétien-Guillaume de Lamoignon de Malesherbes (1721–94), another friend of Turgot and like him a minister of the crown, protected the *encyclopédistes.* On balance, however, it is hard to see how the monarchy, even if it had resolved its financial problems, which it was very far from doing, could have extended this ecumenism from art to politics and social life. To do so, it would have had to transform its institutions in keeping with new conceptions regarding men's public and private affairs and to commit itself to the rejection of the corporatist ethic in economic life. Thus the monarchy seemed fated to failure and the stage set for revolution. (P.Hi.)

The French Revolution and Napoleon, 1789–1815

THE DESTRUCTION OF THE ANCIEN RÉGIME

1789: the convergence of revolutions. *The juridical revolution.* Louis XVI's decision to convene the Estates-General in May 1789 became a turning point in French history. When he invited his subjects to express their opinions and grievances in preparation for this event—unprecedented in living memory—hundreds responded with pamphlets in which the liberal ideology of 1789 gradually began to take shape. Exactly how the Estates-General should deliberate proved to be the pivotal consciousness-raising issue. Each of the three Estates could vote separately (by order) as they had in the distant past, or they could vote jointly (by head). Because the Third Estate was to have twice as many deputies as the others, only voting by head would assure its preponderant influence. If the estates voted by order, the clergy and nobility would effectively exercise a veto power over important decisions. Most pamphleteers of 1789 considered themselves "patriots," or reformers, and (though some were nobles themselves) identified the excessive influence of "aristocrats" as a chief obstacle to reform. In his influential tract *Qu'est-ce que le tiers état?* (1789; *What is the Third Estate?*) the constitutional theorist Emmanuel-Joseph Abbé Sieyès asserted that the Third Estate really was the French nation. While commoners did all the truly laborious and productive work of society, he claimed with some exaggeration, the nobility monopolized its lucrative sinecures and honours. As a condition of genuine reform, the Estates-General would have to change that situation.

A seismic shift was occurring in elite public opinion. What began in 1787–88 as a conflict between royal authority and traditional aristocratic groups had become a triangular struggle, with "the people" opposing both absolutism and privilege. A new kind of political discourse was emerging, and within a year it was to produce an entirely new concept of sovereignty with extremely far-reaching implications.

Patriots were driven to increasingly bold positions in part by the resistance and bad faith of royal and aristocratic forces. It is not surprising that some of the Third Estate's most radical deputies came from Brittany, whose nobility was so hostile to change that it finally boycotted the Estates-General altogether. Hoping that the king would take the lead of the patriot cause, liberals were disappointed at the irresolute, business-as-usual attitude of the monarchy when the Estates opened at Versailles in May 1789. While the nobility organized itself into a separate chamber (by a vote of 141 to 47), as did the clergy (133 to 114), the Third Estate refused to do so. After pleading repeatedly for compromise and debating their course of action in the face of this deadlock, the Third Estate's deputies finally acted decisively. On June 17 they proclaimed that they were not simply the Third Estate of the Estates-General but a National Assembly, which the other deputies were invited to join. A week later 150 deputies of the clergy did indeed join the National Assembly, but the nobility protested that the whole notion was illegal.

Now the king had to clarify his position. He began by closing the hall assigned to the Third Estate and ordering all deputies to hear a royal address on June 23. The deputies, however, adjourned to an indoor tennis court on the 21st and there swore a solemn oath to continue meeting until they had provided France with a constitution. Two days later they listened to the king's program for reform. In the "royal session" of June 23 the king pledged to honour civil liberties, agreed to fiscal equality (already conceded by the nobility in its *cahiers*, or grievance petitions), and promised that the Estates-General would meet regularly in the future. But, he declared, they would deliberate separately by order. France was to become a constitutional monarchy, but one in which "the ancient distinction of the three orders will be conserved in its entirety." In effect the king was forging an alliance with the nobility, who only a year before had sought to hobble him. For the patriots this was too little and too late.

The Tennis Court Oath

In a scene of high drama, the deputies refused to adjourn to their own hall. When ordered to do so by the king's chamberlain, the Assembly's president, astronomer Jean-Sylvain Bailly (1736–93), responded—to the official's amazement—that "the assembled nation cannot receive orders." Such defiance unnerved the king. Backing down, he directed the nobles several days later to join a National Assembly whose existence he had just denied. Thus the Third Estate, with its allies in the clergy and nobility, had apparently effected a successful nonviolent revolution from above. Having been elected in the *bailliages* (the monarchy's judicial districts which served as electoral circumscriptions) to represent particular constituents to their king, the deputies had transformed themselves into representatives of the entire nation. Deeming the nation alone to be sovereign, they as its representatives claimed sole authority to exercise that sovereignty. This was the juridical revolution of 1789.

The juridical revolution

Parisian revolt. In fact, the king had by no means reconciled himself to this revolutionary act. His concession was a strategic retreat until he could muster the military power to subdue the patriots. Between June 27 and July 1 he ordered 20,000 royal troops into the Paris region, ostensibly to protect the assembly and to prevent disorder in the restive capital. The assembly's pleas to the king to withdraw these menacing and unnecessary troops fell on deaf ears. For all their moral force, the deputies utterly lacked material force to counter the king's obvious intentions. The assembly was saved from likely dissolution only by a massive popular mobilization.

During the momentous political events of 1788–89 much of the country lay in the grip of a classic subsistence crisis. Bad weather had reduced the grain crops that year by almost one-quarter the normal yield. An unusually cold winter compounded the problem, as frozen rivers halted the transport and milling of flour in many localities. Amidst fears of hoarding and profiteering, grain and flour reserves dwindled. In Paris the price of the four-pound loaf of bread—the standard item of consumption accounting for most of the population's calories and nutrition—rose from its usual 8 sous to 14 sous by January 1789. This intolerable trend set off traditional forms of popular protest. If royal officials did not assure basic food supplies at affordable prices, then people would act directly to seize food. During the winter and spring of 1789 urban consumers and peasants rioted at bakers and markets and attacked millers and grain convoys. Then, in July, this anxiety merged with the looming political crisis at Versailles. Parisians believed that food shortages and royal troops would be used in tandem to starve the people and overwhelm them into submission. They feared an "aristocratic plot" to throttle the patriot cause.

When the king dismissed the still-popular finance minister Jacques Necker on July 11, Parisians correctly read this as a signal that the counterrevolution was about to begin. Instead of yielding, however, they rose in rebellion. Street-corner orators such as Camille Desmoulins stirred their compatriots to resist. Confronting royal troops in the streets, they won some soldiers over to their side and induced officers to confine other potentially unreliable units to their barracks. On July 13, bands of Parisians ransacked armourers' shops in a frantic search for weapons. The next day a large crowd invaded the Hôtel des Invalides and seized thousands of rifles without resistance. Then they moved to the Bastille, an old fortress commanding the Faubourg Saint-Antoine, which had served as a notorious royal prison earlier in the century but was now scheduled for demolition. Believing that gunpowder was stored there, the crowd laid siege to the Bastille. Unlike the troops at the Invalides, the Bastille's tiny garrison resisted, a fierce battle erupted, and dozens of Parisians were killed. When the garrison finally capitulated, the irate crowd massacred several of the soldiers. In another part of town two leading royal officials were lynched for their presumed role in the plot against the people. Meanwhile the electors of Paris, who had continued to meet after choosing their deputies to the Estates-General, ousted the royal officials of the city government, formed a revolutionary municipality, and organized a citizens' militia or national guard to patrol the streets. Similar municipal revolutions occurred in 26 of the 30 largest French cities, thus assuring that the capital's defiance would not be an isolated act.

The fall of the Bastille

By any standard, the fall of the Bastille to the Parisian crowd was a spectacular symbolic event—a seemingly miraculous triumph of the people against the power of royal arms. The heroism of the crowd and the blood of its martyrs—ordinary Parisian artisans, tradesmen, and workers—sanctified the patriot cause. Most importantly, the elites and the people of Paris had made common cause, despite the inherent distrust and social distance between them. The mythic unity of the Third Estate—endlessly invoked by patriot writers and orators—seemed actually to exist, if only momentarily. Before this awesome material and moral force Louis XVI capitulated. He did not want civil war in the streets. The Parisian insurrection of July 14 not only saved the National Assembly from dissolution but altered the course of the Revolution by giving it a far more active, popular, and violent dimension. On July 17 the king traveled to Paris, where he publicly donned a cockade bearing a new combination of colours: white for the Bourbons and blue and red for the city of Paris. This tricolour was to become the new national flag.

Peasant insurgencies. Peasants in the countryside, meanwhile, carried on their own kind of rebellion, which combined traditional aspirations and anxieties with support of the patriot cause. The peasant revolt was autonomous, yet it reinforced the urban uprising to the benefit of the National Assembly.

Competition over the ownership and the use of land had intensified in many regions. Peasants owned only about 40 percent of the land (see above *Agricultural patterns*), leasing or sharecropping the rest from the nobility, the urban middle class, and the church. Population growth and subdivision of the land from generation to generation was reducing the margin of subsistence for many families. Innovations in estate management—the grouping of leaseholds, conversion of arable land to pasture, enclosure of open fields, division of common land at the lord's initiative, discovery of new seigneurial dues or arrears in old ones—exasperated peasant tenants and smallholders. Historians debate whether these were capitalistic innovations or traditional varieties of seigneurial extraction, but in either case the countryside was boiling with discontent over these trends as well as over oppressive royal taxes and food shortages. Peasants were poised between great hopes for the future raised by the calling of the Estates-General and extreme anxiety—fear of losing land, fear of hunger (especially after the catastrophic harvest of 1788), and fear of a vengeful and mighty aristocracy.

In July peasants in several regions sacked the castles of nobles and burned the documents that recorded their feudal obligations. This peasant insurgency eventually merged into the movement known as the Great Fear. Rumours abounded that these vagrants were actually brigands in the pay of nobles, who were marching on villages to destroy the new harvest and coerce the peasants into submission. The fear was baseless, but hundreds of false alarms and panics stirred up hatred and suspicion of nobles, led peasants to

The Great Fear

arm themselves as best they could, and set off widespread attacks on chateaus and feudal documents. The peasant revolt suggested that the unity of the Third Estate against "aristocrats" extended from Paris to villages across the country. The Third Estate truly seemed invincible.

The abolition of feudalism. Of course the violence of peasant insurgency worried the deputies of the National Assembly; to some it seemed as if the countryside were being engulfed by anarchy that threatened all property. But the majority was unwilling to turn against the rebellious peasants. Instead of denouncing their violence, they tried to appease peasant opinion. Liberal nobles and clergy began the session of August 4th by renouncing their ancient feudal privileges. Within hours the Assembly was propelled into decreeing "the abolition of feudalism" as well as the church tithe, venality of office, regional privilege, and fiscal privilege. A few days later, to be sure, the Assembly clarified the August 4th decree to assure that "legitimate" seigneurial property rights were maintained. While personal feudal servitudes such as hunting rights, seigneurial justice, and labour services were suppressed outright, most seigneurial dues were to be abolished only if the peasants paid compensation to their lords, set at 20 to 25 times the annual value of the obligation. The vast majority of peasants rejected that requirement by passive resistance, until pressure built in 1792–93 for the complete abolition of all seigneurial dues without compensation.

The abolition of feudalism was crucial to the evolution of a modern, contractual notion of property and to the development of an unimpeded market in land. But it did not directly affect the ownership of land or the level of ordinary rents and leases. Seigneurs lost certain kinds of traditional income, but they remained landowners and landlords. While all peasants gained in dignity and status, only the landowning peasants came out substantially ahead economically. Tenant farmers found that what they had once paid for the tithe was added on to their rent. And the Assembly did virtually nothing to assure better lease terms for renters and sharecroppers, let alone their acquisition of the land they tilled.

Declaration of the Rights of Man and of the Citizen

The new regime. By sweeping away the old web of privileges, the August 4th decree permitted the Assembly to construct a new regime. Since it would take months to draft a constitution, the Assembly on August 27 promulgated its basic principles in a Declaration of the Rights of Man and of the Citizen. A rallying point for the future, the declaration also stood as the death certificate of the ancien régime. The declaration's authors believed it to have universal significance. "In the new hemisphere, the brave inhabitants of Philadelphia have given the example of a people who reestablished their liberty," conceded one deputy, but "France would give that example to the rest of the world." At the same time the declaration responded to particular circumstances and was thus a calculated mixture of general principles and specific concerns. Its concept of natural rights meant that the Revolution would not be bound by history and tradition but could reshape the contours of society according to reason—a position vehemently denounced by Edmund Burke in England.

The very first article of the declaration resoundingly challenged Europe's old order by affirming that "men are born and remain free and equal in rights. Social distinctions may be based only on common utility." Most of its articles concerned individual liberty, but the declaration's emphasis fell equally on the prerogatives of the state as expressed through law. (Considering how drastically the erstwhile delegates to the Estates-General had exceeded their mandates, they certainly needed to underscore the legitimacy of their new government and its laws.) The declaration, and subsequent revolutionary constitutions, channeled the sovereignty of the nation into representative government, thereby negating claims by *parlements,* provincial estates, or divine-right monarchs as well as any conception of direct democracy. Though the declaration affirmed the separation of powers, by making no provision for a supreme court, it effectively left the French legislature as the ultimate judge of its own actions. The declaration defined liberty as "the ability to do whatever does not harm another . . . whose limits can only be determined by law." The same limitation by positive law was attached to specific liberties, such as freedom from arbitrary arrest, freedom of expression, and freedom of religious conscience. The men of 1789 believed deeply in these liberties, yet they did not establish them in autonomous, absolute terms that would insure their sanctity under any circumstances.

The constitution

Restructuring France. From 1789 to 1791 the National Assembly acted as a constituent assembly, drafting a constitution for the new regime while also governing from day to day. The constitution established a limited monarchy, with a clear separation of powers in which the king was to name and dismiss his ministers. But sovereignty effectively resided in the legislative branch, to consist of a single house, the Legislative Assembly, elected by a system of indirect voting. ("The people or the nation can have only one voice, that of the national legislature," wrote Sieyès. "The people can speak and act only through its representatives.") Besides failing to win a bicameral system, the moderate Anglophile, or *monarchien,* faction lost a bitter debate on the king's veto power: the Assembly granted the king only a suspensive or delaying veto over legislation; if a bill passed the Legislative Assembly in three successive years, it would become law even without royal approval.

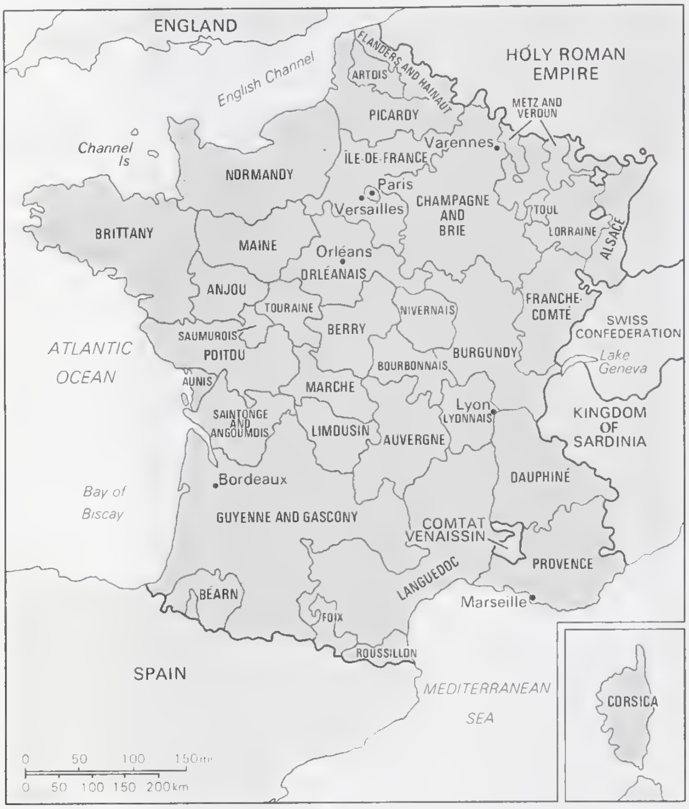
Dismayed at what he deemed the ill-considered radicalism of such decisions, Jean-Joseph Mounier, a leading patriot deputy in the summer of 1789 and author of the Tennis Court Oath, resigned from the Assembly in October. In a similar vein, some contemporary historians (notably François Furet) have suggested that the Assembly's integral concept of national sovereignty and legislative supremacy effectively reestablished absolutism in a new guise, providing the new government with inherently unlimited powers. Nor, they believe, is it surprising that the revolutionaries abused those powers as their pursuit of utopian goals encountered resistance. In theory this may well be true, but it must be balanced against the actual institutions created to implement those powers and the spirit in which they were used. With a few exceptions—notably the religious issue—the National Assembly acted in a liberal spirit, more pragmatic than utopian, and was decidedly more constructive than repressive.

The revolutionaries took civil equality seriously but created a limited definition of political rights. They effectively transferred political power from the monarchy and the privileged estates to the general body of propertied citizens. Nobles lost their privileges in 1789 and their titles in 1790, but as propertied individuals they could readily join the new political elite. The constitution restricted the franchise to "active" citizens who paid a minimal sum in taxes, with higher property qualifications for eligibility for public office (a direct tax payment equivalent to three days' wages for voting and 10 days' wages for electors and officeholders). Under this system about two-thirds of adult males had the right to vote for electors and to choose certain local officials directly. Although it favoured wealthier citizens, the system was vastly more democratic than Britain's.

Women's rights and roles

Predictably, the franchise did not extend to women, despite delegations and pamphlets advocating women's rights. The Assembly responded brusquely that, because women were too emotional and easily misled, they must be kept out of public life and devote themselves to their nurturing and maternal roles. But the formal exclusion of women from politics did not keep them on the sidelines. Women were active combatants in local conflicts that soon erupted over religious policy, and they agitated over subsistence issues—Parisian women, for example, made a mass march to Versailles in October that forced the king to move back to the capital. In the towns, they formed auxiliaries to local Jacobin clubs and even a handful of independent women's clubs, participated in civic festivals, and did public relief work.

The Assembly's design for local government and administration proved to be one of the Revolution's most durable legacies. Obliterating the political identity of France's historic provinces, the deputies redivided the nation's territory into 83 *départements* of roughly equal size. Unlike the old provinces, each department would have exactly



Gouvernements before 1789.

the same institutions; departments were in turn subdivided into districts, cantons, and communes (the common designation for a village or town). On the one hand, this administrative transformation promoted decentralization and local autonomy: citizens of each department, district, and commune elected their own local officials. On the other hand, these local governments were subordinated to the national legislature and ministries in Paris. The departments therefore became instruments of national uniformity and integration, which is to say, centralization. This ambiguity the legislators fully appreciated, assuming that a healthy equilibrium could be maintained between the two tendencies. That the revolutionary government of 1793 and Napoleon later used these structures to concentrate power from the centre was not something they could anticipate.

The new administrative map also created the parameters for judicial reform. Sweeping away the entire judicial system of the ancien régime, the revolutionaries established a civil court in each district and a criminal court in each department. At the grass roots they replaced seigneurial justice with a justice of the peace in each canton. Judges on all these tribunals were to be elected. While rejecting the use of juries in civil cases, the Assembly decreed that felonies would be tried by juries; if a jury convicted, judges would merely apply the mandatory sentences set out in the Assembly's tough new penal code of 1791. Criminal defendants also gained the right to counsel, which had been denied them under the jurisprudence of the ancien régime. In civil law, the Assembly encouraged arbitration and mediation to avoid the time-consuming and costly processes of formal litigation. In general, the revolutionaries hoped to make the administration of justice more accessible and expeditious.

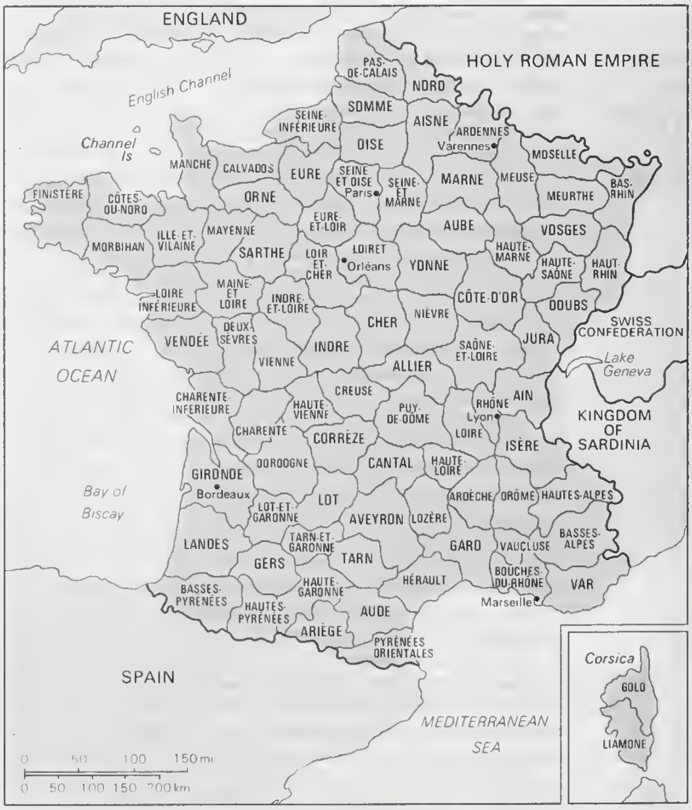
Economic reforms

Guided by laissez-faire doctrine and its hostility to privileged corporations, the Assembly sought to open up economic life to unimpeded individual initiative and competition. Besides proclaiming the right of all citizens to enter any trade and conduct it as they saw fit, the Assembly dismantled internal tariffs and chartered trading monopolies and abolished the guilds of merchants and artisans. Insisting that workers must bargain in the economic marketplace as individuals, the Le Chapelier Law of June 1791 banned workers' associations and strikes. The precepts of economic individualism extended to rural life as well. In theory, peasants and landlords were now free to cultivate their fields as they wished, regardless of traditional collective routines and constraints. In practice, however, communal restraints proved to be deep-rooted and resistant to legal abolition.

Sale of national lands. The Assembly had not lost sight of the financial crisis that precipitated the collapse of absolutism in the first place. Creating an entirely new option for its solution, the Assembly voted to place church property—about 10 percent of the land in France—"at the disposition of the nation." This property was designated as *biens nationaux,* or national lands. The government then issued large-denomination notes called assignats, underwritten and guaranteed by the value of that land. It intended to sell off national lands to the public, which would pay for it in assignats that would then be retired. Thus church property would in effect pay off the national debt and obviate the need for further loans. Unfortunately the temptation to print additional assignats proved too great. Within a year the assignat evolved into a paper currency in small and large denominations, with sharp inflationary effects.

Sale of national lands

As the national lands went on sale, fiscal needs took priority over social policy. Sales were arranged in large lots and at auction in the district capitals—procedures that favoured wealthier buyers. True, for about a year in 1793–94, after émigré property was added to the *biens nationaux,* large lots were divided into small parcels. In addition, small peasants acquired some of this land through resale by the original buyers. But overall the urban middle classes and large-scale peasants emerged with the bulk of this land, to the intense frustration of small peasants. The French historian Georges Lefebvre's study of the Nord department, for example, found that 7,500 bourgeois purchased 48 percent of the land, while 20,300 peasants bought 52 percent. But the top 10 percent of these peasant purchasers accounted for 60 percent of the peasants' total. Whatever the social origins of the buyers, however, they were likely to be reliable supporters of the Revolution if only to guarantee the security of their new acquisitions.



Revolutionary departments after 1789.

Seeds of discord. Security could not be taken for granted, however, because the Revolution progressively alienated or disappointed important elements of French society. Among the elites, opposition began almost immediately when some of the king's close relatives left the country in disgust after July 14, thus becoming the first émigrés. Each turning point in the Revolution touched off new waves of emigration, especially among the nobility. By 1792 an estimated two-thirds of the royal officer corps had resigned their commissions and most had left the country. A contentious royalist press bitterly denounced the policies of the Assembly as spoliation and the revolutionary atmosphere as a form of anarchy. Abroad, widespread enthusiasm for the events in France among the general public from London to Vienna was matched by intense hostility in ruling circles, fearful of revolutionary contagion within their own borders.

After the first months of solidarity, long-standing urban-rural tensions took on new force. Though peasants might vote in large numbers, the urban middle classes predictably emerged with the lion's share of the new district and departmental offices after the first elections of 1790. Administrative and judicial reform gave these local officials more powers for intrusion into rural society than royal officials ever had, with battalions of armed national guards to back them up. Peasants might easily view urban revolutionary elites as battening on political power and national lands. And, while the Assembly made the tax system more uniform and equitable, direct taxes remained heavy and in formerly privileged regions actually rose, while nothing was done to relieve the plight of tenant farmers. Later, when the revolutionary government sought to draft young men into the army, another grievance was added to the list.

The Civil Constitution of the Clergy

Religious tensions. But it was religious policy that most divided French society and generated opposition to the Revolution. Most priests had initially hoped that sweeping reform might return Roman Catholicism to its basic ideals, shorn of aristocratic trappings and superfluous privileges, but they assumed that the church itself would collaborate in the process. In the Assembly's view, however, nationalization of church property gave the state responsibility for regulating the church's temporal affairs, such as salaries, jurisdictional boundaries, and modes of clerical appointment. On its own authority the Assembly reduced the number of dioceses and realigned their boundaries to coincide with the new departments, while requesting local authorities to redraw parish boundaries in conformity with population patterns. Under its Civil Constitution of the Clergy (July 1790) bishops were to be elected by departmental electoral assemblies, while parish priests were to be chosen by electors in the districts. Clerical spokesmen deplored the notion of lay authority in such matters and insisted that the Assembly must negotiate reforms with a national church council.

In November 1790 the Assembly forced the issue by requiring all sitting bishops and priests to take an oath of submission. Those who refused would lose their posts, be pensioned off, and replaced by the prescribed procedures. Throughout France a mere seven bishops complied, while only 54 percent of the parish clergy took the oath. Contrary to the Assembly's hopes, the clergy had split in two, with "constitutional" priests on one side and "refractory" priests on the other. Regional patterns accentuated this division: in the west of France, where clerical density was unusually high, only 15 percent of the clergy complied.

The schism quickly engulfed the laity. As refractories and constitutionals vied for popular support against their rivals, parishioners could not remain neutral. Intense local discord erupted over the implementation of the Civil Constitution of the Clergy. District administrations backed by urban national guards intervened to install "outsiders" chosen to replace familiar or even beloved refractory priests in many parishes; villagers responded by badgering or boycotting the hapless priests who took the oath. Opinion on both sides tended to fateful extremes, linking either the Revolution with impiety or the Roman Catholic church with counterrevolution.

Political tensions. The political life of the new regime was also proving more contentious than the revolutionaries had anticipated. With courage and consistency, the Assembly had provided that officials of all kinds be elected. But it was uncertain whether these officials, once the ballots were cast, could do their duty free from public pressure and agitation. Nor was it clear what the role of "public opinion" and the mechanisms for its expression would be. The spectacular development of a free press and political clubs provided an answer. Fearful that these extraparliamentary institutions could be abused by demagogues, the Assembly tried to curb them from time to time but to no avail. Freed entirely from royal censorship, writers and publishers rushed to satisfy the appetite for news and political opinion. The first journalists included deputies reporting to their constituents by means of a newspaper. Paris, which had only four quasi-official newspapers at the start of 1789, saw more than 130 new periodicals by the end of the year, most admittedly short-lived, including 20 dailies. As the journalist Jacques-Pierre Brissot put it, newspapers are "the only way of educating a large nation unaccustomed to freedom or to reading, yet looking to free itself from ignorance." Provincial publishers were as quick to found new periodicals in the larger towns. Bordeaux, for example, had only one newspaper in 1789, but 16 appeared within the next two years. While some papers remained bland and politically neutral, many had strong political opinions.

Revolutionary clubs

Like the National Assembly, revolutionary clubs also began at Versailles, when patriot deputies rallied to a caucus of outspoken Third Estate deputies from Brittany. This Breton Club, complete with by-laws, minutes, committees, correspondence, and membership requirements, began to call itself "The Society of the Friends of the Constitution." Soon it was known as the Jacobin Club, after the Dominican convent where the club met when the assembly transferred to Paris in October. Most prominent revolutionaries belonged to the Jacobin Club, from constitutional royalists such as Mirabeau, Lafayette, and Barnave to radicals like Brissot, Pétion, and Robespierre. By mid-1791, however, moderates became uncomfortable at the Jacobin Club, where Maximilien Robespierre was emerging as a dominant figure.

The Jacobin Club was pushed from the left by the Cordeliers Club, one of the neighbourhood clubs in the capital. The Cordeliers militants rejected the Assembly's concept of representation as the exclusive expression of popular sovereignty. They held to a more direct vision of popular sovereignty as relentless vigilance and participation by citizens through demonstrations, petitions, deputations, and if necessary insurrection. In his newspaper *L'Ami du peuple* ("The Friend of the People") Jean-Paul Marat injected an extreme rhetoric about alleged conspiracies and the need for violence against counterrevolutionaries that exceeded anything heard in the Assembly's political discourse.

Like the press, clubs quickly spread in the provinces. Building no doubt on old-regime patterns of sociability—reading clubs, Freemasonry, or confraternities—political clubs became a prime vehicle for participation in the Revolution. More than 300 towns had clubs by the end of 1790, and 900 by mid-1791. Later clubs spread to the villages as well: a study has counted 5,000 localities that had clubs at one time or another between 1790 and 1795. Many clubs affiliated with the Paris Jacobin Club, the "mother club," in an informal nationwide network. Most began with membership limited to the middle class and a sprinkling of liberal nobles, but gradually artisans, shopkeepers, and peasants joined the rolls. Initially the clubs promoted civic education and publicized the Assembly's reforms. But some became more activist, seeking to influence political decisions with petitions, to exercise surveillance over constituted authorities, and to denounce those they deemed remiss.

By 1791 the assembly found itself in a cross-fire between the machinations of counterrevolutionaries—émigrés, royalist newspapers, refractory clergy—and the denunciations of radicals. Its ability to steer a stable course depended in part on the cooperation of the king. Publicly Louis XVI distanced himself from his émigré relatives, but privately he was in league with them and secretly corresponded with the royal houses of Spain and Austria to enlist their support. On June 21, 1791, the royal family attempted to flee its "captivity" in the Tuileries Palace and escape across the Belgian border. Rashly, Louis left behind a letter revealing his utter hostility to the Revolution. At the last minute, however, the king was recognized at the town of Varennes near the border, and the royal party was forcibly returned to Paris.

Splitting of the Jacobin Club

A great crisis for the Revolution ensued. While the Assembly reinforced the frontiers by calling for 100,000 volunteers from the national guard, its moderate leaders hoped that this fiasco would end Louis's opposition once and for all. In order to preserve their constitutional compromise, they turned a blind eye to the king's manifest treason by inventing the fiction that he had been kidnapped. As Antoine Barnave put it: "Are we or are we not going to terminate the Revolution? Or are we going to start it all over again?" Outside the Assembly, however, Jacobins and Cordeliers launched a petition campaign against reinstating the king. A mass demonstration on July 17 at the Champ de Mars against the king ended in a bloody riot, as the authorities called out the national guard under Lafayette's command to disperse the demonstrators. This precipitated vehement recriminations in the Jacobin Club, which finally split apart under the pressure. The mass of moderate deputies abandoned the club to a rump of radicals and formed a new association called the Feuillant Club. Under the leadership of Robespierre and Jérôme Pétion (who later became mortal enemies), the purged Jacobin Club rallied most provincial clubs and emerged from the crisis with a more unified, radical point of view. For the time being, however, the moderates prevailed in the Assembly. They completed the Constitution of 1791, and on the last day of September 1791 the National Assembly dissolved itself, having previously decreed the ineligibility of its members for the new Legislative Assembly.

When the newly elected Legislative Assembly convened in October, the question of counterrevolution dominated its proceedings. Jacobin deputies like Brissot argued that only war against the émigré army gathering at Coblenz across the Rhine could end the threat: "Do you wish at one blow to destroy the aristocracy, the refractory priests, and the malcontents: then destroy Coblenz." Whereas the Feuillants opposed this war fever, Lafayette saw a successful military campaign as a way to gain power, while the king's circle believed that war would bring military defeat to France and a restoration of royal authority. On the other side, the Habsburg monarch, Leopold II, had resisted the pleas of his sister Marie-Antoinette and opposed intervention against France, but his death in March 1792 brought his bellicose son Francis II to the throne and the stage was set for war.

The Brunswick Manifesto

In April 1792 France went to war against a coalition of Austria, Prussia, and the émigrés. Each camp expected rapid victory, but both were disappointed. The allies repulsed a French offensive and soon invaded French territory. The Legislative Assembly called for a new levy of 100,000 military volunteers, but, when it voted to incarcerate refractory clergy, the king vetoed the decree. Though many Frenchmen remained respectful of the king, the most vocal elements of public opinion denounced Louis and demonstrated against him; but the Legislative Assembly refused to act. As Prussian forces drove toward Paris, their commander, the Duke of Brunswick, proclaimed his aim of restoring the full authority of the monarchy and warned that any action against the king would bring down "exemplary and memorable vengeance" against the capital. Far from terrifying the Parisians, the Brunswick Manifesto enraged them and drove them into decisive action.

Militants in the Paris Commune, the revolutionary government of Paris set up by the capital's 48 wards or sections, gave the Legislative Assembly a deadline in which to suspend the king. When it passed unheeded, they organized an insurrection. On Aug. 10, 1792, a huge crowd of armed Parisians stormed the royal palace after a fierce battle with the garrison. The Legislative Assembly then had no choice but to declare the king suspended. That night more than half the deputies themselves fled Paris, for the Legislative Assembly too had lost its mandate. Those who remained ordered the election by universal male suffrage of a National Convention. It would judge the king, draft a new republican constitution, and govern France during the emergency. The Constitution of 1791 had lasted less than a year, and the second revolution dreaded by the Feuillants had begun.

THE FIRST FRENCH REPUBLIC

The second revolution. The insurrection of Aug. 10, 1792, did not of course stop the Prussian advance on the capital. As enthusiastic contingents of volunteers left for the front, fear of counterrevolutionary plots gripped the capital. Journalists like Jean-Paul Marat pointed to the prisons bursting with vagrants and criminals as well as refractory clergy and royalists and asked what would happen if traitors forced open the jails and released these hordes of fanatics and brigands. In response Parisians took the law into their own hands with an orgy of mass lynching.

Prison massacres

On their own initiative, citizens entered the prisons, set up "popular tribunals" to hold perfunctory trials, and summarily executed between 1,100 and 1,400 prisoners out of a total of 2,800, stabbing and hacking them to death with any instruments at hand. These prison massacres were no momentary fit of frenzy but went on for four days. At the time no one in authority dared try to stop the slaughter. Officials of the provisional government and the Paris Commune "drew a veil" over this appalling event as it ran its course, though soon political rivals were accusing each other of instigating the massacres. In a different vein, Robespierre among others concluded that popular demands for vengeance and terror had to be channeled into legal forms; to prevent such anarchy, the state itself must become the orderly instrument of the people's punitive will.

The next two weeks brought this period of extreme uncertainty to a close. On September 20 the French army turned back the invaders at the Battle of Valmy, and in November at the Battle of Jemappes it won control of the Austrian Netherlands (now Belgium). On September 21 the National Convention convened, ending the vacuum of authority that had followed the August 10 insurrection. Its first major task was to decide the fate of the ex-king. The Convention's trial of Louis became an educational experience for the French people in which the institution of monarchy was to be completely desacralized.

Execution of Louis Capet

Hard evidence of Louis's treason produced a unanimous guilty verdict, but the issue of punishment divided the deputies sharply. In a painstaking and solemn debate each deputy cast his vote individually and explained it. At the end the Convention voted the death sentence, 387 to 334. A motion for reprieve was defeated (380 to 310), and one to submit the verdict to a national referendum was rejected (425 to 286). This ill-considered proposal left the impression that certain deputies were frantic to save the king's life, and their Jacobin opponents were quick to raise vague accusations of treasonous intent against them. In any event, the former King Louis XVI, now known simply as "Citizen Capet," was executed on Jan. 21, 1793, in an act of immense symbolic importance. For the deputies to the National Convention, now regicides, there could be no turning back. Laws to deport the refractory clergy, to bar the émigrés forever upon pain of death, and to confiscate their property rounded out the Convention's program for eliminating the Revolution's most determined enemies.

A republic in crisis. By the spring of 1793, however, the republic was beleaguered. In the second round of the war, the coalition—now reinforced by Spain, Piedmont, and Britain—routed French forces in the Austrian Netherlands and the Rhineland and breached the Pyrenees. Fighting on five different fronts and bereft of effective leadership, French armies seemed to be losing everywhere. Even General Charles-François Dumouriez, the hero of the first Netherlands campaign, had gone over to the enemy in April after quarreling with the Convention. Meanwhile, civil war had broken out within France. Rural disaffection in western France, especially over the religious question referred to earlier, had been building steadily, leaving republicans in the region's cities and small towns an unpopular and vulnerable minority. Rural rage finally erupted into armed rebellion when in March 1793 the Convention decreed that each department must produce a quota of citizens for the army. In four departments south of the Loire River the Vendée rebellion began with assaults on the towns and the massacre of patriots. Gradually royalist nobles assumed the leadership of the peasants and weavers who had risen on their own initiative. Forging them into a "Catholic and Royalist Army," they hoped to overthrow the republic and restore the Bourbons.

The Convention could take no comfort from the economic situation either. An accelerating depreciation of the assignats compounded severe shortages of grain and flour in 1793. Inflation, scarcity, and hoarding made life unbearable for the urban masses and hampered efforts to provision the republic's armies. In reaction to such economic hardships and to the advance of antirepublican forces at the frontiers and within France, Parisian radicals clamoured relentlessly for decisive action such as price controls and the repression of counterrevolutionaries.

Girondins and Montagnards. The Convention, however, was bitterly divided almost to the point of paralysis. From the opening day, two outspoken groups of deputies vied for the support of their less factional colleagues. The roots of this rivalry lay in a conflict between Robespierre and Brissot for leadership of the Jacobin Club in the spring and summer of 1792. At that time Robespierre had argued almost alone against the war that Brissot passionately advocated. Later, when the war went badly and the Brissotins, anxious to wield executive power, acted equivocally in their relations with the king, the Jacobins turned on them. Brissot was formally expelled from the club in October, but his expulsion merely formalized a division that had already crystallized during the elections to the Convention in the previous month.

The Paris electoral assembly sent Robespierre, Marat, Georges-Jacques Danton, and other stalwarts of the Paris Commune and the Jacobin Club to the Convention, while systematically rejecting Brissot and his allies such as the former mayor of Paris, Pétion. The Parisian deputies and their provincial supporters, numbering between 200 and 300 (depending on which historian's taxonomy one accepts), took seats on the Convention's upper benches and came to be known as the Montagnards, or the Mountain.

Supported by a network of journalists and by politicians such as Interior Minister Jean-Marie Roland, however, the Brissotins retained their popularity in the provinces and were returned as deputies by other departments. In the Convention the Brissotin group included most deputies from the department of the Gironde, and the group came to be known by their opponents as the Girondins. The inner core of this loose faction, who often socialized in Jeanne-Marie Roland's salon, numbered about 60, and with their supporters perhaps 150 to 175.

Differences between Girondins and Montagnards

At bottom the Girondin-Montagnard conflict stemmed from a clash of personalities and ambitions. Over the years, historians have made the case for each side by arguing that their opponents constituted the truly aggressive or obstructive minority seeking to dominate the Convention. Clearly most deputies were put off by the bitter personal attacks that regularly intruded on their deliberations. The two factions differed most over the role of Paris and the best way to deal with popular demands. Though of similar middle-class background as their rivals, the Montagnards sympathized more readily with the sansculottes (the local activists) of the capital and proved temperamentally bolder in their response to economic, military, and political problems. United by an extreme hostility to Parisian militance, the Girondins never forgave the Paris Commune for its inquisitorial activity after August 10. Indeed, some Girondins did not feel physically secure in the capital. They also appeared more committed to political and economic liberties and therefore less willing to adopt extreme revolutionary measures no matter how dire the circumstances. Ready to set aside similar constitutional scruples, the Montagnards tailored their policies to the imperatives of "revolutionary necessity" and unity.

While the Girondins repeatedly attacked Parisian militants—at one point demanding the dissolution of the Paris Commune and the arrest of its leaders—the Montagnards gradually forged an informal alliance with the sansculottes. Similarly, the Montagnards supported deputies sent on mission to the departments when they clashed with locally elected officials, while the Girondins tended to back the officials. The Montagnards therefore alienated many moderate republicans in the provinces. As deputies of the centre, or "Plain," such as Bertrand Barère, vainly tried to mediate between the two sides, the convention navigated through this factionalism as best it could and improvised new responses to the crisis: a revolutionary tribunal to try political crimes; local surveillance committees to seek out subversives; and a Committee of Public Safety to coordinate measures of revolutionary defense. By the end of May 1793 a majority seemed ready to support the Mountain.

Believing that the Girondins had betrayed and endangered the republic, the Paris sections (with the connivance of the Montagnards and the Paris Jacobin Club) demanded in petitions that the Convention expel the "perfidious deputies." On May 31 they mounted a mass demonstration and on June 2 forced a showdown by deploying armed national guards around the convention's hall. Backed by a huge crowd of unarmed men and women, their solid phalanx of fixed bayonets made it impossible for the deputies to leave without risking serious violence. Inside, the Montagnards applauded this insurrection as an expression of popular sovereignty, akin to that of July 14 or August 10. When the people thus spoke directly, they argued, its deputies had no choice but to comply. Centrists did everything they could to avoid a purge but in the end decided that only this fateful act could preserve the Revolution's unity. Barère composed a report to the French people justifying the expulsion of 29 Girondins. Later 120 deputies who signed a protest against the purge were themselves suspended from the Convention, and in October the original Girondins stood trial before the revolutionary tribunal, which sentenced them to death. The Montagnard ascendancy had begun.

Though the deadlock in the Convention was now broken, the balance of forces in the country was by no means clear. The Parisian sansculottes might well continue to intimidate the convention and emerge as the dominant partner in their alliance with the Montagnards—just as Girondin orators had warned. Conversely, provincial opinion might rebel against this mutilation of the National Convention by Paris and its Montagnard partisans. Purged of the Girondins, the Convention itself could reach consensus more readily, but the nation as a whole was more divided than ever.

At first it seemed as if the expulsion of the Girondins would indeed backfire. More than half of the departmental directories protested against the Convention's purge. But, faced with the pleas for unity and the threats from the Convention, most of this opposition subsided quickly. Only 13 departments continued their defiant stance, and only 6 of these passed into overt armed rebellion against the Convention's authority. Still, this was a serious threat in a country already beleaguered by civil war and military reversals. The Jacobins stigmatized this new opposition as the heresy of federalism—implying that the "federalists" no longer believed in a unified republic. Jacobin propaganda depicted the federalists as counterrevolutionaries. In fact, most were moderate republicans hostile to the royalists and committed to constitutional liberties. They did not intend to overthrow the republic or separate from it. Rather they hoped to wrest power back from what they deemed the tyrannical alliance of Montagnards and Parisian sansculottes.

Uprisings in Lyon and Marseille

In Lyon, Marseille, Toulon, and Bordeaux, bitter conflicts between local moderates and Jacobins contributed decisively to the rebellion. Uprisings in Lyon and Marseille (France's second- and third-largest cities) began in late May when moderates seized power from local Jacobin authorities who had threatened their lives and property—Jacobins like the firebrand Marie-Joseph Chalier in Lyon who was supported by Montagnard representatives-on-mission. The expulsion of the Girondins was merely the last straw. Whatever its causes, however, "federalist" rebellion did threaten national unity and the Convention's sovereign authority. Royalists, moreover, did gain control of the movement in Toulon and opened that port to the British. Holding out no offer of negotiation, the Convention organized military force to crush the rebellions and promised the leaders exemplary punishment. "Lyon has made war against liberty," declared the Convention; "Lyon no longer exists." When the republic's forces recaptured the city in October, they changed its name to "Liberated City," demolished the houses of the wealthy, and summarily executed more than 2,000 Lyonnais, including many wealthy merchants.

The Reign of Terror. After their victory in expelling the Girondins, Parisian militants "regenerated" their own sectional assemblies by purging local moderates, while radicals like Jacques-René Hébert and Pierre-Gaspard Chaumette tightened their grip on the Paris Commune. On Sept. 5, 1793, they mounted another mass demonstration to demand that the Convention assure food at affordable prices and "place terror on the order of the day." Led by its Committee of Public Safety, the Convention placated the popular movement with decisive actions. It proclaimed the need for terror against the Revolution's enemies, made economic crimes such as hoarding into capital offenses, and decreed a system of price and wage controls known as the maximum. The Law of Suspects empowered local revolutionary committees to arrest "those who by their conduct, relations or language spoken or written, have shown themselves partisans of tyranny or federalism and enemies of liberty." In 1793–94 well over 200,000 citizens were detained under this law; though most of them never stood trial, they languished in pestiferous jails, where an estimated 10,000 perished. About 17,000 death sentences were handed down by the military commissions and revolutionary tribunals of the Terror, 72 percent for charges of armed rebellion in the two major zones of civil war—the federalist southeast and the western Vendée region. One-third of the departments, however, had fewer than 10 death sentences passed on their inhabitants and were relatively tranquil.

The maximum

To help police the maximum and requisition grain in the countryside, as well as to carry out arrest warrants and guard political prisoners, the Convention authorized local authorities to create paramilitary forces. About 50 such *armées révolutionnaires* came into being as ambulatory instruments of the Terror in the provinces. Fraternizing with peasants and artisans in the hinterland, these forces helped raise revolutionary enthusiasm but ultimately left such village sansculottes vulnerable to the wrath of the wealthy citizens whom they harassed.

Back in June the Convention had quickly drafted a new democratic constitution, incorporating such popular demands as universal male suffrage, the right to subsistence, and the right to free public education. In a referendum this Jacobin constitution of 1793 was approved virtually without dissent by about two million voters. Because of the emergency, however, the Convention placed the new constitution on the shelf in October and declared that "the provisional government of France is revolutionary until the peace." There would be no elections, no local autonomy, no guarantees of individual liberties for the duration of the emergency. The Convention would rule with a sovereignty more absolute than the old monarchy had ever claimed. Nor would serious popular protest be tolerated any longer, now that the Jacobins had used such intervention to secure power. The balance in the alliance between Montagnards and sansculottes gradually shifted from the streets of Paris to the halls and committee rooms of the Convention.

The Jacobin constitution

From the beginning a popular terrorist mentality had helped shape the Revolution. Peasants and townspeople alike had been galvanized by fear and rage over "aristocratic plots" in 1789. Lynchings of "enemies of the people" punctuated the Revolution, culminating in the September massacres, which reflected an extreme fear of betrayal and an unbridled punitive will. Now the Revolution's leaders were preempting this punitive will in order to control it: they conceived of terror as rational rather than emotional and as organized rather than instinctive. Paradoxically they were trying to render terror lawful—legality being an article of faith among most revolutionaries—but without the procedural safeguards that accompanied the regular criminal code of 1791.

For the more pragmatic Montagnards that deviation was justified by the unparalleled emergency situation confronting France in 1793: before the benefits of the Revolution could be enjoyed, they must be secured against their enemies by force. ("Terror is nothing other than justice, prompt, severe, inflexible . . . . Is force made only to protect crime?" declared Robespierre.) For the more ideologically exalted Jacobins like Robespierre and Saint-Just, however, the terror would also regenerate the nation by promoting equality and the public interest. In their minds a link existed between terror and virtue: "virtue, without which terror is fatal; terror, without which virtue is powerless." Whoever could claim to speak for the interests of the people held the mantle of virtue and the power of revolutionary terror.

The Jacobin dictatorship. The Convention consolidated its revolutionary government in the Law of 14 Frimaire Year II (Dec. 4, 1793; for a discussion of the Revolutionary calendar, see below). To organize the Revolution, to promote confidence and compliance, efficiency and control, this law centralized authority in a parliamentary dictatorship, with the Committee of Public Safety at the helm. The committee already controlled military policy and patronage; henceforth local administrators (renamed national agents), tribunals, and revolutionary committees also came under its scrutiny and control. The network of Jacobin clubs was enlisted to monitor local officials, nominate new appointees, and in general to serve as "arsenals of public opinion."

The Committee of Public Safety

Opposed to "ultrarevolutionary" behaviour and uncoordinated actions even by its own deputies-on-mission, the committee tried to stop the dechristianization campaigns that had erupted during the anarchic phase of the Terror in the fall of 1793. Usually instigated by radical deputies, the dechristianizers vandalized churches or closed them down altogether, intimidated constitutional priests into resigning their vocation, and often pressured them into marrying to demonstrate the sincerity of their conversion. Favouring a diestic form of civil religion, Robespierre implied that the atheism displayed by some dechristianizers was a variant of counterrevolution. He insisted that citizens must be left free to practice the Roman Catholic religion, though for the time being most priests were not holding services.

The committee also felt strong enough a few months later to curb the activism of the Paris sections, dissolve the *armées révolutionnaires,* and purge the Paris Commune—ironically what the Girondins had hoped to do months before. But in this atmosphere no serious dissent to official policy was tolerated. The once vibrant free press had been muzzled after the purge of the Girondins. In March 1794 Hébert and other "ultrarevolutionaries" were arrested, sent to the revolutionary tribunal, and guillotined. A month later Danton and other so-called "indulgents" met the same fate for seeking to end the Terror—prematurely in the eyes of the committee. Then the Convention passed the infamous law of 22 Prairial Year II (June 10, 1794) to streamline revolutionary justice, denying the accused any effective right to self-defense and eliminating all sentences other than acquittal or death. Indictments by the public prosecutor, now virtually tantamount to a death sentence, multiplied rapidly.

The Terror was being escalated just when danger no longer threatened the republic—after French armies had prevailed against Austria at the decisive Battle of Fleurus on June 26 and long after rebel forces in the Vendée, Lyon, and elsewhere had been vanquished. By June 1794 the Jacobin dictatorship had forged an effective government and had mobilized the nation's resources, thereby mastering the crisis that had brought it into being. Yet, on 8 Thermidor, Robespierre took the rostrum to proclaim his own probity and to denounce yet another unnamed group as traitors hatching "a conspiracy against liberty." Robespierre had clearly lost his grip on reality in his obsession with national unity and virtue. An awkward coalition of moderates, Jacobin pragmatists, rival deputies, and extremists who rightly felt threatened by the Incorruptible

Fall of Robespierre

(as he was known) finally combined to topple Robespierre and his closest followers. On 9 Thermidor (July 27, 1794) the Convention ordered the arrest of Robespierre and Saint-Just, and, after a failed resistance by loyalists in the Paris Commune, they were guillotined without trial the following day. The Terror was over.

The Army of the Republic. The Jacobin dictatorship had been an unstable blend of exalted patriotism, resolute political leadership, ideological fanaticism, and populist initiatives. The rhetoric and symbolism of democracy constituted a new civic pedagogy, matched by bold egalitarian policies. The army was a primary focal point of this democratic impetus. Back in 1790 the National Assembly had opted for a small military of long-term professionals. One-year volunteers bolstered the line army after the outbreak of war, and in March 1793 the convention called for an additional 300,000 soldiers, with quotas to be provided by each department. Finally, in August 1793 it decreed the *lévee en masse*—a "requisition" of all able-bodied, unmarried men between the ages of 18 and 25. Despite massive draft evasion and desertion, within a year almost three-quarters of a million men were under arms, the citizen-soldiers merged with line-army troops in new units called demibrigades. This huge popular mobilization reinforced the revolution's militant spirit. The citizen-soldiers risking their lives at the front had to be supported by any and all means back home, including forced loans on the rich and punitive vigilance against those suspected of disloyalty.

Within the constraints of military discipline, the army became a model of democratic practice. Both noncommissioned and commissioned officers were chosen by a combination of election and appointment, in which seniority received some consideration, but demonstrated talent on the battlefield brought the most rapid promotion. The republic insisted that officers be respectful toward their men and share their privations. Jacobin military prosecutors enforced the laws against insubordination and desertion but took great pains to explain them to the soldiers and to make allowances for momentary weakness in deciding cases. Soldiers received revolutionary newspapers and sang revolutionary songs, exalting the citizen-soldier as the model sansculotte. Meanwhile, needy parents, wives, or dependents of soldiers at the front received subsidies, while common soldiers seriously wounded in action earned extremely generous veterans' benefits.

The Revolution's egalitarian promise never involved an assault on private property, but its concept of "social limitations" on property made it possible for the Convention to abolish all seigneurial dues without compensation, abolish slavery in the colonies (where slave rebellions had already achieved that result in practice), endorse the idea of progressive taxation, and temporarily regulate the economy in favour of consumers. In 1793–94 the Convention enacted an unprecedented national system of public assistance entitlements, with one program allocating small pensions to poor families with dependent children and another providing pensions to aged and indigent farm workers, artisans, and rural widows—the neediest of the needy. "We must put an end to the servitude of the most basic needs, the slavery of misery, that most hideous of inequalities," declared Bertrand Barère of the Committee of Public Safety. The Convention also implemented the Revolution's long-standing commitment to primary education with a system of free public primary schooling for both boys and girls. The Lakanal Law of November 1794 authorized public schools in every commune with more than 1,000 inhabitants, the teachers to be selected by examination and paid fixed salaries by the government.

The Thermidorian reaction. With control passing from the Montagnards after Robespierre's fall, moderates in the convention hoped to put the Terror and sansculotte militance behind them while standing fast against counterevolution and rallying all patriots around the original principles of the Revolution. But far from stabilizing the Revolution, the fall of "the tyrant" on 9 Thermidor set in motion a brutal struggle for power. Those who had suffered under the Terror now clamoured for retribution, and moderation quickly gave way to reaction. As federalists were released, Jacobins were arrested; as the suspended Girondins were reinstated, Montagnards were purged; as moderates could feel safe, Jacobins and sansculottes were threatened. Like the Terror, the Thermidorian reaction had an uncontrollable momentum of its own. Antiterrorism—in the press, the theatre, the streets—degenerated into a "white terror" against the men of the Year II. In the south, especially in Provence and the Rhône valley, the frontier between private feuds and political reaction blurred as law and order broke down. Accounts were settled by lynchings, murder gangs, and prison massacres of arrested sansculottes.

"White terror"

Alongside this political reaction, Thermidor set off a new economic and monetary crisis. Committed to free-trade principles, the Thermidorians dismantled the economic regulation and price controls of the Year II, along with the apparatus of the Terror that had put teeth into that system. The depreciation of the assignats, which the Terror had halted, quickly resumed. By 1795 the cities were desperately short of grain and flour, while meat, fuel, dairy products, and soap were entirely beyond the reach of ordinary consumers. By the spring of 1795 scarcity was turning into famine for working people of the capital and other cities. Surviving cadres of sansculottes in the Paris sections mobilized to halt the reaction and the economic catastrophe it had unleashed. After trying petitions and demonstrations, a crowd of sansculottes invaded the Convention on 1 Prairial Year III (May 20, 1795) in what was to prove the last popular uprising of the French Revolution. "Bread is the goal of their insurrection, physically speaking," reported a police observer, "but the Constitution of 1793 is its soul." This rear-guard rebellion of despair was doomed to fail, despite the support of a few remaining Montagnard deputies, whose fraternization with the demonstrators was to cost them their lives after the insurgents were routed the following day.

Instead of implementing the democratic Constitution of 1793, the Thermidorian Convention was preparing a new, more conservative charter. Anti-Jacobin and antiroyalist, the Thermidorians clung to the elusive centre of the political spectrum. Their constitution of 1795 (Year III) established a liberal republic with a franchise based on the payment of taxes similar to that of 1791, a two-house legislature to slow down the legislative process, and a five-man executive Directory to be chosen by the legislature. Within a liberal framework, the central government retained great power, including emergency powers to curb freedom of the press and freedom of association. Departmental and municipal administrators were to be elected but could be removed by the Directory, and commissioners appointed by the Directory were to monitor them and report on their compliance with the laws.

The Thermidorian constitution

The Directory. The new regime, referred to as the Directory, began auspiciously in October 1795 with a successful constitutional plebiscite and a general amnesty for political prisoners. But as one of its final acts the Convention added the "Two-thirds Decree" to the package, requiring for the sake of continuity that two-thirds of its deputies must sit by right in the new legislature regardless of voting in the departments. This outraged conservatives and royalists hoping to regain power legally, but their armed uprising in Paris was easily suppressed by the army. The Directory also weathered a conspiracy on the far left by a cabal of unreconciled militants organized around a program of communistic equality and revolutionary dictatorship. The Babeuf plot was exposed in May 1796 by a police spy, and a lengthy trial ensued in which François-Noël ("Gracchus") Babeuf, the self-styled "Tribune of the People," was sentenced to death.

Apart from these conspiracies, the political life of the Directory revolved around annual elections to replace one-third of the deputies and local administrators. The spirit of the "Two-thirds Decree" haunted this process, however, since the directors believed that stability required their continuation in power and the exclusion of royalists or Jacobins. The Directory would tolerate no organized opposition. During or immediately after each election, the government in effect violated the constitution in order to save it, whenever the right or the left seemed to be gaining ground.

Elections and coups

As a legacy of the nation's revolutionary upheavals, elections under the Directory displayed an unhealthy combination of massive apathy and rancorous partisanship by small minorities. When the elections of 1797 produced a royalist resurgence, the government responded with the coup of Fructidor Year V (September 1797), ousting two of the current directors, arresting leading royalist politicians, annulling the elections in 49 departments, shutting down the royalist press, and resuming the vigorous pursuit of returned émigrés and refractory clergy. This heartened the Neo-Jacobins, who organized new clubs called "constitutional circles" to emphasize their adherence to the regime. But this independent political activism on the left raised the spectre of 1793 for the Directory, and in turn it closed down the Neo-Jacobin clubs and newspapers, warned citizens against voting for "anarchists" in the elections of 1798, and promoted schisms in electoral assemblies when voters spurned this advice. When democrats (or Neo-Jacobins) prevailed nonetheless, the Directory organized another purge in the coup of Floréal Year VI (May 1798) by annulling all or some elections in 29 departments. Ambivalent and faint-hearted in its republican commitment, the Directory was eroding political liberty from within. But as long as the Constitution of 1795 endured, it remained possible that political liberty and free elections might one day take root.

Sister republics. Meanwhile the Directory regime successfully exported revolution abroad by helping to create "sister republics" in western Europe. During the Revolution's most radical phase, in 1793–94, French expansion had stopped more or less at the nation's self-proclaimed "natural frontiers"—the Rhine, Alps, and Pyrenees. The Austrian Netherlands (now Belgium) and the left bank of the Rhine had been major battlefields in the war against the coalition, and French victories in those sectors were followed by military occupation, requisitions, and taxation, but also by the abolition of feudalism and similar reforms. In 1795 both areas were annexed to France, and their territories were divided into departments, which would henceforth be treated like other French departments.

Strategic considerations and French national interest were the main engines of French foreign policy in the revolutionary decade but not the only ones. Elsewhere in Europe native patriots invited French support against their own ruling princes or oligarchies. As the historian R.R. Palmer has argued, Europe was divided not simply by a conflict between Revolutionary France and other states but by conflicts within various states between revolutionary or democratic forces and conservative or traditional forces. Indeed, abortive revolutionary movements had already occurred in the Austrian Netherlands and in the United Provinces (Dutch Netherlands). The ideals of liberty, equality, or popular sovereignty knew no borders.

By 1797 Prussia and Spain had made peace with France, but Austria and Britain continued the struggle. In a new strategy the French launched an attack across the Alps aimed at Habsburg Lombardy, from which they hoped to drive north toward Vienna. Commanded by General Napoleon Bonaparte, this campaign succeeded beyond expectations. In the process northern Italy was liberated from Austria, and the Habsburgs were driven to the peace table, where they signed the Treaty of Campo Formio on Oct. 17, 1797. An invasion of the Netherlands, home base of British forces on the Continent, produced a similar victory. In short order two "sister republics" were proclaimed by native revolutionaries under French protection—the Cisalpine Republic in northern Italy and the Batavian Republic (formerly the Dutch Netherlands). These were later joined by the Helvetic Republic in Switzerland, and two very shaky republics—the Roman Republic in central Italy and the Parthenopean Republic in the south around Naples. All these republics were exploited financially by the French, but then again their survival depended on the costly presence of French troops. The French interfered in their internal politics, but this was no more than the Directory was doing at home. Because these republics could not defend themselves in isolation, they acted as sponges on French resources as much as they provided treasure or other benefits to France. France's extended lines of occupation made it extremely vulnerable to attack when Britain organized a second coalition in 1798 that included Russia and Austria. But, when the battles were over, Switzerland, northern Italy, and the Netherlands remained in the French sphere of influence.

The treasure coming from the sister republics was desperately needed in Paris since French finances were in total disarray. The collapse of the assignats and the hyperinflation of 1795–96 not only destroyed such social programs as public assistance pensions and free public schooling but also strained the regime's capacity to keep its basic institutions running. In 1797 the government finally engineered a painful return to hard currency and in effect wrote down the accumulated national debt by two-thirds of its value in exchange for guaranteeing the integrity of the remaining third.

Alienation and coups. After the Fructidor coup of 1797 the Directory imprudently resumed the republic's assault on the Roman Catholic religion. Besides prohibiting the outward signs of Catholicism such as the ringing of church bells or the display of crosses, the government revived the Revolutionary calendar. Instituted in 1793, the new calendar featured a 10-day week called the *décade,* designed to swallow up the Christian Sunday in a new cycle of work and recreation. While it had fallen into disuse after the Thermidorian reaction, the Directory ordered in 1798 that the *décadi* be treated as the official day of rest for workers and businesses as well as public employees and schoolchildren. Forbidding organized recreation on Sundays, the regime also pressured Catholic priests to celebrate mass on the *décadi* rather than on ex-Sundays. This aggressive confrontation with the habits and beliefs of most French citizens sapped whatever shreds of popularity the regime still had.

The revolutionary calendar

French citizens were already alienated by the Directory's foreign policy and its new conscription law. Conscription became a permanent obligation of young men between the ages of 20 and 25 under the Loi Jourdan of Sept. 5, 1798. To fight the War of the Second Coalition that began in 1799, the Directory mobilized three "classes," or age cohorts, of young men but encountered massive draft resistance and desertion in many regions. Meanwhile, retreating armies in the field lacked rations and supplies because, it was alleged, corrupt military contractors operated in collusion with government officials. This war crisis prompted the legislature to oust four of the directors (the Prairial coup of June 18, 1799) and allowed a brief resurgence of Neo-Jacobin agitation for drastic emergency measures.

In reality the balance of power was swinging toward a group of disaffected conservatives. Led by Sieyès, one of the new directors, these "revisionists" wished to escape from the instability of the Directory regime, especially its tumultuous annual elections and its cumbersome separation of powers. They wanted a more reliable structure of political power, which would allow the new elite to govern securely and thereby guarantee the basic reforms and property rights of 1789. Ironically, the Neo-Jacobins, or democrats, stood as the constitution's most ardent defenders against the maneuvers of these "oligarchs."

Using mendacious allegations about Neo-Jacobin plots as a cover, the revisionists prepared a parliamentary coup to jettison the constitution. To provide the necessary military insurance, the plotters sought a leading general. Though he was not their first choice, they eventually enlisted Napoleon Bonaparte—recently returned from his Egyptian campaign, about whose disasters the public knew almost nothing. Given a central role in the coup, which occurred on the 18th Brumaire Year VIII (Nov. 9, 1799), General Bonaparte addressed the legislature, and when some deputies balked at his call for scrapping the constitution, his troopers cleared the hall. A rump of each house then convened to draft a new constitution, and during these deliberations Bonaparte shouldered aside Sieyès and emerged as the dominant figure in the new regime. Brumaire was not really a military coup and did not at first produce a dictatorship. It was a parliamentary coup to create a new constitution and was welcomed by people of differing opinions who saw in it what they wished to see.

The Brumaire coup

The image of an energetic military hero impatient with the abuses of the past must have seemed reassuring.

THE NAPOLEONIC ERA

The Consulate. The "revisionists" who engineered the Brumaire coup intended to create a strong, elitist government that would curb the republic's political turmoil and guarantee the conquests of 1789. They had in mind what might be called a senatorial oligarchy rather than a personal dictatorship. General Bonaparte, however, advocated a more drastic concentration of power. Within days of the coup, Bonaparte emerged as the dominant figure, an insistent and persuasive presence who inspired confidence. Clearly outmaneuvered, Sieyès soon withdrew from the scene, taking with him his complex notions of checks and balances. While the regime, known as the Consulate, maintained a republican form, Bonaparte became from its inception a new kind of authoritarian leader.

Emperor Napoleon

Approved almost unanimously in a plebiscite by three million votes (of which half may have been manufactured out of thin air), the Constitution of the Year VIII created an executive consisting of three consuls, but the First Consul wielded all real power. That office was of course vested in Napoleon Bonaparte. In 1802, after a string of military and diplomatic victories, another plebiscite endowed him with the position for life. By 1804 Bonaparte's grip on power was complete, and belief in his indispensability was pervasive in the governing class. In April 1804 various government bodies agreed "that Napoleon Bonaparte be declared Emperor and that the imperial dignity be declared hereditary in his family." The Constitution of the Year XII (May 1804) establishing the empire was approved in a plebiscite by more than 3,500,000 votes against about 8,000. (After this point General Bonaparte was known officially as Napoleon.)

The Constitution of 1791, the Convention, and the Directory alike had been organized around representation and legislative supremacy, the fundamental political principles first proclaimed in June 1789 by the National Assembly. This tradition came to an end with the Consulate. Its new bicameral legislature lost the power to initiate legislation; now the executive branch drafted new laws. One house (the Tribunate) debated such proposals, either endorsed or opposed them, and then sent deputies to present its opinion to the other house, the Legislative Corps (Corps Législatif), which also heard from government spokesmen. Without the right to debate, the Legislative Corps then voted on whether to enact the bill. Even these limited powers were rarely used independently, since both houses were appointed in the first instance by the government and later renewed by cooption. When certain tribunes such as Benjamin Constant did manifest a critical spirit, they were eventually purged, and in 1807 the Tribunate was suppressed altogether. On the whole, then, the legislative branch of government became little more than a rubber stamp.

The Council of State

After Brumaire, Sieyès had envisaged an independent institution called the Senate to conserve the constitution by interpreting it in the light of changing circumstances. In practice, the Senate became the handmaiden of Bonaparte's expanding authority, sanctioning changes such as the life consulship, the purge of the Tribunate, and Napoleon's elevation to the rank of hereditary emperor. For creating "legislation above the laws" at Bonaparte's behest, its 80 handpicked members were opulently rewarded with money and honours. As power shifted decisively to the executive branch, Bonaparte enlisted a new institution called the Council of State to formulate policy, draft legislation, and supervise the ministries. Appointed by the First Consul, this body of experienced jurists and legislators was drawn from across the former political spectrum. Talent and loyalty to the new government were the relevant criteria for these coveted posts.

The Consulate did not entirely eliminate the electoral principle from the new regime, but voters were left with no real power, and elections became an elaborate charade. Citizens voted only for electoral colleges, which in turn created lists of candidates from which the government might fill occasional vacancies in the Legislature or Senate. In the event, the primary assemblies of voters were rarely convened, and membership in the electoral colleges became a kind of honorific lifetime position. The judiciary too lost its elective status. In the hope of creating a more professional and compliant judiciary, the Consulate's sweeping judicial reform provided for lifetime appointments of judges—which did not prevent Napoleon from purging the judiciary in 1808. Napoleon was also disposed to eliminate criminal juries as well, but the Council of State prevailed on him to maintain them.

The prefects

Successive Revolutionary regimes had always balanced local elections with central control, but the Consulate destroyed that balance completely. The Local Government Act of February 1800 eliminated elections for local office entirely and organized local administration from the top down. To run each department, the Consulate appointed a prefect, reminiscent of the old royal intendants, who was assisted by subprefects on the level of the arrondissements (subdistricts of the departments) and by appointed mayors in each commune. The original Revolutionary commitment to local autonomy gave way before the rival principles of centralization and uniformity. The prefect became the cornerstone of the Napoleonic dictatorship, supervising local government at all levels, keeping careful watch over his department's "public spirit," and above all assuring that taxes and conscripts flowed in smoothly. While even the most trivial local matter had to be referred to the prefect, all major decisions taken by the prefect had in turn to be sanctioned by the interior ministry in Paris.

Loss of political freedom. Politics during the Directory had been marked by an unwholesome combination of ferocious partisanship and massive apathy. Weary of political turmoil and disillusioned by politicians of all kinds, most Frenchmen now accepted the disappearance of political freedom and participation with equanimity. The few who still cared passionately enough to resist collided with the apparatus of a police state. A regime that entirely avoided genuine elections would scarcely permit open political dissent. Where the Directory had been ambivalent about freedom of association, for example, the Consulate simply banned political clubs outright and placed Jacobin and royalist cadres under surveillance by the police ministry. In 1801, blaming democratic militants for a botched attempt to assassinate him with a bomb as his carriage drove down the Rue Saint-Nicaise—a plot actually hatched by fanatical royalists—Bonaparte ordered the arrest and deportation to Guiana of about 100 former Jacobin and sansculotte militants. In 1804 he had the Duke d'Enghien, a member of the Bourbon family, abducted from abroad, convicted of conspiracy by a court-martial, and executed.

Reimposition of censorship

Outspoken liberals also felt the lash of Bonaparte's intolerance for any kind of opposition. After he purged the Tribunate, the consul registered his displeasure with the salon politics of liberal intellectuals by dissolving the Class of Moral and Political Science of the National Institute in 1803. One of the most principled liberals, Madame de Stäel, chose to go into exile rather than exercise the self-censorship demanded by the regime. Meanwhile, the only newspapers tolerated were heavily censored. Paris, for example, had more than 70 newspapers at the time of the Brumaire coup; by 1811 only four quasi-official newspapers survived, ironically the same number as had existed before 1789. In the provinces each department had at most one newspaper, likewise of quasi-official character. The reimposition of censorship was matched by Napoleon's astute management of news and propaganda.

Society in Napoleonic France. *Religious policy.* If the Consulate's motto was "authority from above, confidence from below," Bonaparte's religious policy helped secure that confidence. The concordat negotiated with the papacy in 1802 reintegrated the Roman Catholic church into French society and ended the cycle of bare toleration and persecution that had begun in 1792. Having immediately halted the campaign to enforce the republican calendar (which was quietly abolished in 1806), the Consulate then extended an olive branch to the refractory clergy. The state continued to respect the religious freedom of non-Catholics, but the concordat recognized Catholicism as "the preferred religion" of France—in effect, though not in name, the nation's established religion. Upkeep of the church became a significant item in local budgets, and the clergy regained de facto control over primary education. The state, however, retained the upper hand in church-state relations. By signing the concordat, the pope accepted the nationalization of church property in France and its sale as *biens nationaux*. Bishops, though again consecrated by Rome, were named by the head of state, and the government retained the right to police public worship.

The most conservative Catholics looked askance at the concordat, which in their eyes promoted an excessively national or Gallican church rather than a truly Roman Catholic church. They correctly suspected that Bonaparte—personally a religious skeptic—would use it as a tool of his own ambitions. Indeed, he declared that the clergy would be his "moral prefects," propagating traditional values and obedience to authority. Later, for example, the clergy was asked to teach an imperial catechism, which would "bind the consciences of the young to the august person of the Emperor."

Calvinists and Jews

The Napoleonic regime also organized France's approximately one million Calvinists into hierarchical "consistories" subject to oversight by the state. Protestant pastors, paid by the state, were designated by the elders who led local congregations and consistories; the more democratic tendencies of Calvinism were thus weakened in exchange for official recognition. France's 60,000 Jews, residing mainly in Alsace and Lorraine, were also organized into consistories. Like priests and pastors, their rabbis were enlisted to promote obedience to the laws, though they were not salaried by the state. Official recognition, however, did not prevent discriminatory measures against Jews. A law of 1808, ostensibly for "the social reformation of the Jews," appeased peasant debtors in Alsace by canceling their debts to Jewish moneylenders.

Napoleonic nobility. Napoleon cultivated the loyalty of the nation's wealthy landed proprietors by a system of patronage and honours. He thereby facilitated the emergence of a ruling class drawn from both the middle classes and the nobility of the old regime, which had been divided by the artificial barriers of old-regime estates and privileges. The principal artifacts of Napoleon's social policy were the lists he ordered of the 600 largest taxpayers in each department, most having incomes of at least 3,000 livres a year. Inclusion on these lists became an insignia of one's informal status as a notable. Members of the electoral colleges and departmental advisory councils were drawn from these lists. Although such honorific positions had little power and no privileges, the designees were effectively co-opted into the regime. Napoleon's Legion of Honour, meanwhile, conferred recognition on men who served the state, primarily military officers who largely stood outside the ranks of the landed notables. By 1814 the Legion had 32,000 members, of whom only 1,500 were civilians.

The new nobility

After Napoleon had himself crowned emperor in 1804, he felt the need for a court aristocracy that would lend lustre and credibility to his new image. He also reasoned that only by creating a new nobility based on merit could he displace and absorb the old nobility, which had lost its titles in 1790 but not its pretensions. By 1808 a new hierarchy of titles had been created, which were to be hereditary provided that a family could support its title with a large annual income—30,000 livres, for example, in the case of a count of the empire. To facilitate this, the emperor bestowed huge landed estates and pensions on his highest dignitaries. The Napoleonic nobility, in other words, would be a veritable upper class based on a combination of service and wealth. Predictably, the new nobility was top-heavy with generals (59 percent altogether), but it also included many senators, archbishops, and members of the Council of State; 23 percent of the Napoleonic nobility were former nobles of the ancien régime. These social innovations endured after Napoleon's fall—the Bourbons adopted the system of high-status electoral colleges, maintained the Legion of Honour, and even allowed the Napoleonic nobles to retain their titles alongside the restored old-regime nobility.

The Civil Code. The Napoleonic Civil Code, however, had a far greater impact on postrevolutionary society than did the social innovations. This ambitious work of legal codification, perhaps the crowning glory of the Council of State, consolidated certain basic principles established in 1789: civil equality and equality before the law; the abolition of feudalism in favour of modern contractual forms of property; and the secularization of civil relations. Codification also made it easier to export those principles beyond the borders of France. In the area of family relations, however, the Napoleonic Code was less a codification of revolutionary innovations than a reaction against them. By reverting to patriarchal standards that strengthened the prerogatives of the husband and father, it wiped out important gains that women had made during the Revolution. The code's spirit on this subject was summed up in its statement that "a husband owes protection to his wife; a wife owes obedience to her husband." Wives were again barred from signing contracts without their husbands' consent, and a wife's portion of the family's community property fell completely under her husband's control during his lifetime. The code also curbed the right of equal inheritance, which the Revolution had extended even to illegitimate children, and increased the father's disciplinary control over his children.

Divorce legislation

The code also rolled back the Revolution's extremely liberal divorce legislation. When marriage became a civil rite rather than an obligatory religious sacrament in 1792, divorce became possible for the first time. Divorce could be obtained by mutual consent but also for a range of causes including desertion and simple incompatibility. Under the Napoleonic Code contested divorce was possible only for unusually cruel treatment resulting in grave injury and for adultery on the part of the wife. Faced with an unfaithful husband, however, "the wife may demand divorce on the ground of adultery by her husband [only] when he shall have brought his concubine into their common residence."

Napoleonic policy frequently reacted against the Revolution's liberal individualism. While the regime did not restore the guilds outright, for example, it reimposed restrictive or even monopolistic state regulation on such occupational groups as publishers and booksellers, the Parisian building trades, attorneys, barristers, notaries, and doctors. Napoleon wished to strengthen the ties that bound individuals together, which derived from religion, the family, and state authority. Napoleon's domestic innovations—the prefectorial system, with its extreme centralization of administrative authority; the University, a centralized educational bureaucracy that scrutinized all types of teachers; the concordat with the Vatican that reversed the secularizing tendencies of the Revolution; the Civil Code, which strengthened property rights and patriarchal authority; and the Legion of Honour, which rewarded service to the state—all endured in the 19th century despite a succession of political upheavals. Historians who admire Napoleon consider these innovations as the "granite masses" on which modern French society developed.

Campaigns and conquests: 1797 to 1807. Napoleon's sway over France depended from the start on his success in war. After his conquest of northern Italy in 1797 and the dissolution of the first coalition, General Bonaparte intended to invade Britain, France's century-long rival and the last remaining belligerent. Concluding that French naval power could not sustain a seaborne invasion, however, he launched a military expedition to Egypt instead, hoping to choke off the main route to Britain's Indian empire. When the expedition bogged down in disease and military stalemate, its commander quietly slipped past a British naval blockade to return to France, where (in the absence of accurate news from Egypt) he was received as the nation's leading military hero.

At the time of the Brumaire coup the republic's armies had been driven from Italy by a second coalition, but they had halted a multifront assault on France by the armies of Russia, Austria, and Britain. The republic, in other words, was no longer in imminent military danger, but the prospect of an interminable war loomed on the horizon. After Brumaire the nation expected its new leader to achieve peace through decisive military victory. This promise Bonaparte fulfilled, once again leading

French armies into northern Italy and defeating Austria at the Battle of Marengo in June 1800. Subsequent defeats in Germany drove Austria to sign the peace treaty of Lunéville in February 1801. Deprived of its continental allies for the second time, a war-weary Britain finally decided to negotiate. In March 1802 France and Britain signed the Treaty of Amiens, and for the first time in 10 years Europe was at peace.

Treaty of Amiens

Within two years, however, the two rivals were again in a state of war. Most historians agree that neither imperial power was solely responsible for the breakdown of this peace since neither would renounce its ambitions for supremacy. Napoleon repeatedly violated the treaty's spirit—by annexing Piedmont, occupying the Batavian Republic, and assuming the presidency of the Cisalpine Republic. To Britain, the balance of power in Europe required an independent Italy and Dutch Netherlands. Britain violated the letter of the treaty, however, by failing to evacuate the island of Malta as it had promised.

Once again, British naval power frustrated Napoleon's attempt to take the war directly to British soil. At the Battle of Trafalgar (Oct. 21, 1805), British naval gunners decimated the French and Spanish fleets. Against Britain's newly enlisted continental allies Napoleon had better luck, as he confronted them one by one before they could effectively unite. Moving his army rapidly across Europe, Napoleon surprised the Austrians at Ulm and then smashed them decisively at the Battle of Austerlitz (Dec. 2, 1805), probably his most brilliant tactical feat. Under the Treaty of Pressburg (criticized by the French foreign minister Talleyrand as entirely too harsh), Austria paid a heavy indemnity, ceded its provinces of Venetia and Tyrol, and allowed Napoleon to abolish the Holy Roman Empire. Prussia, kept neutral for a time by vague promises of sovereignty over Hanover, finally mobilized against France, only to suffer humiliating defeats at the battles of Jena and Auerstadt in October 1806. The French occupied Berlin, levied a huge indemnity on Prussia, seized various provinces, and turned northern Germany into a French sphere of influence. The ensuing campaign against Russia's army in Europe resulted in a bloody stalemate at the Battle of Eylau (Feb. 8, 1807), leaving Napoleon in precarious straits with extremely vulnerable lines of supply. But, when fighting resumed that spring, the French prevailed at the Battle of Friedland (June 14, 1807), and Tsar Alexander I sued for peace. The Treaty of Tilsit, negotiated by the two emperors, divided Europe into two zones of influence, with Napoleon pledging to aid the Russians against their Ottoman rivals, and Alexander promising to cooperate against Britain.

Napoleon's satellite kingdoms

The Grand Empire. Napoleon now had a free hand to reorganize Europe and numerous relatives to install on the thrones of his satellite kingdoms. The result was known as the Grand Empire. Having annexed Tuscany, Piedmont, Genoa, and the Rhineland directly into France, Napoleon placed the Kingdom of Holland (which until 1806 was the Batavian Commonwealth) under his brother Louis, the Kingdom of Westphalia under his brother Jérôme, the Kingdom of Italy under his stepson Eugène as his viceroy, the Kingdom of Spain under his brother Joseph, and the Grand Duchy of Warsaw (carved out of Prussian Poland) under the nominal sovereignty of his ally the king of Saxony. To link his allied states in northern and southern Germany, Napoleon created the Confederation of the Rhine. Even Austria seemed to fall into Napoleon's sphere of influence with his marriage to Archduchess Maria Louise in 1810. (Since the emperor had no natural heirs from his marriage to Joséphine Beauharnais, he reluctantly divorced Joséphine and in 1810 married the Austrian princess, who duly bore him a son the following year.)

The continental system. Britain, however, was insulated from French military power; only an indirect strategy of economic warfare remained possible. Thus far Britain had driven most French merchant shipping from the high seas, and in desperation French merchants sold most of their ships to neutrals, allowing the United States to surpass France in the size of its merchant fleet. But after his string of military victories Napoleon believed that he could choke off British commerce by closing the Continent to its ships and products. With limited opportunities to sell its manufactured goods, he believed, the British economy would suffer from overproduction and unemployment, while the lack of foreign gold in payment for British exports would bankrupt the treasury. As France moved into Britain's foreign markets, Britain's economic crisis would drive its government to seek peace. Accordingly, Napoleon launched the "continental system": in the Berlin Decree of November 1806, he prohibited British trade with all countries under French influence, including British products carried by neutral shipping. When the British retaliated by requiring all neutral ships to stop at British ports for inspection and licenses, Napoleon threatened to seize any ship stopping at English ports. Thus a total naval war against neutrals erupted.

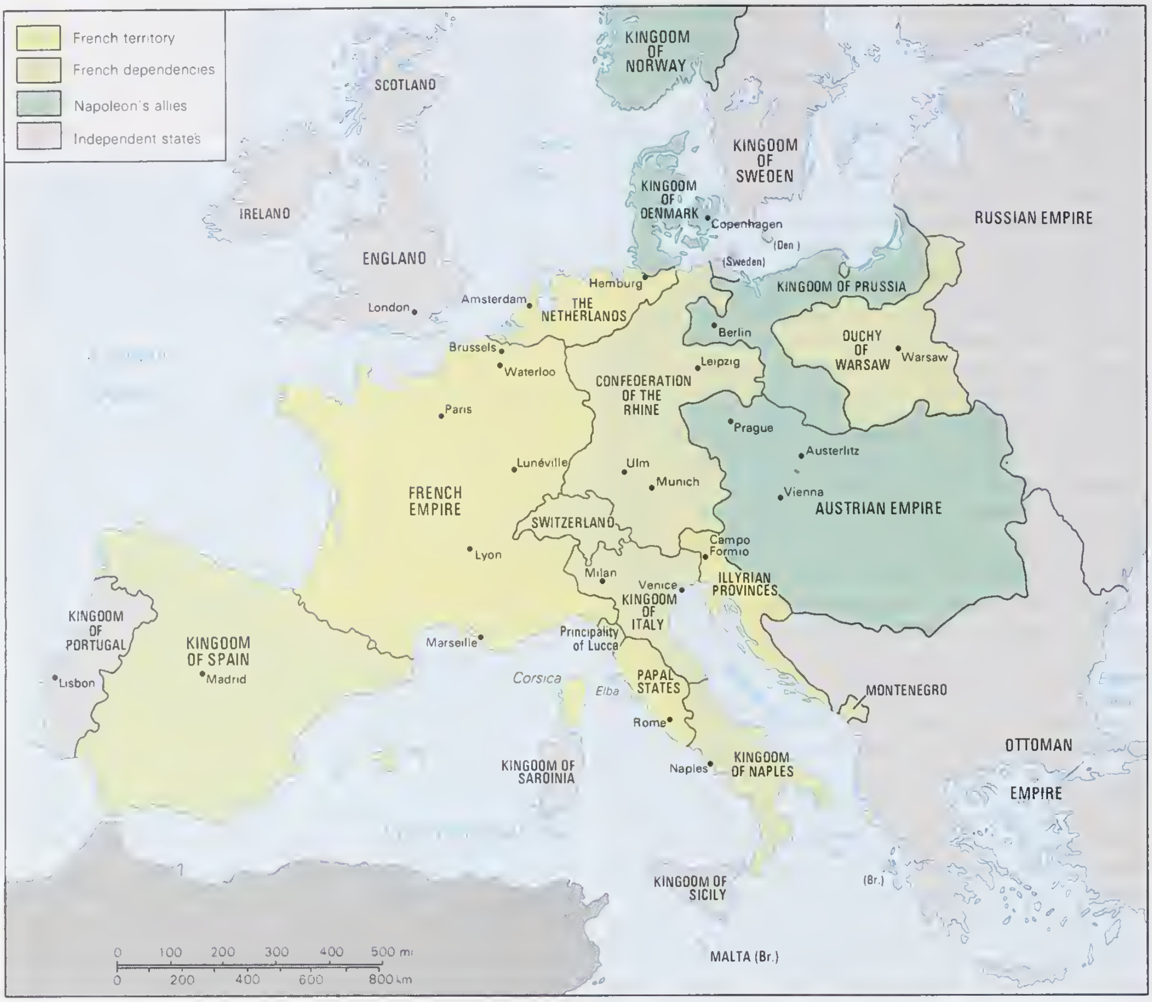
Economic warfare took its toll on all sides. While France did make inroads in cotton manufacturing in the absence of British competition, France and especially its allies suffered terribly from the British blockade, in particular from a dearth of colonial raw materials. The great Atlantic ports of Nantes, Bordeaux, and Amsterdam never recovered, as ancillary industries like shipbuilding and sugar refining collapsed. The axis of European trade shifted decisively inland. The continental system did strain the British economy, driving down exports and gold reserves in 1810, but the blockade was extremely porous. Since Europeans liked British goods, smugglers had incentive to evade the restrictions in places like Spain and Portugal. By 1811, moreover, a restive Tsar Alexander withdrew from the continental system. Thus, the most dire effect of the continental system was the stimulus it gave Napoleon for a new round of aggressions against Portugal, Spain, and Russia.

The invasion of Russia

By 1810 almost 300,000 imperial troops were bogged down in Iberia, struggling against a surprisingly vigorous Spanish resistance and a British expeditionary force. Then, in 1812, Napoleon embarked on his most quixotic aggression—an invasion of Russia designed to humble "the colossus of Northern barbarism" and exclude Russia from any influence in Europe. The Grand Army of 600,000 men that crossed into Russia reached Moscow without inflicting a decisive defeat on the Russian armies. By the time Napoleon on October 19 belatedly ordered a retreat from Moscow, which had been burned to the ground and was unfit for winter quarters, he had lost two-thirds of his troops from disease, battle casualties, cold, and hunger. The punishing retreat through the Russian winter killed most of the others. Yet this unparalleled disaster did not humble or discourage the emperor. Napoleon believed that he could hold his empire together and defeat yet another anti-French coalition that was forming. He correctly assumed that he could still rely on his well-honed administrative bureaucracy to replace the decimated Grand Army.

Conscription. Building on the Directory's conscription law of September 1798, the Napoleonic regime, after considerable trial and error, had created the mechanisms for imposing on the citizens of France and the annexed territories the distasteful obligation of military service. Each year the Ministry of War Administration assigned a quota of conscripts for every department. Using communal birth registers, the mayor of each commune compiled a list of men reaching the age of 19 that year. After a preliminary examination to screen out the manifestly unfit and those below the minimum height of five feet one inch, the young men drew numbers in a lottery at the cantonal seat. Doctors in the departmental capitals later ruled on other claims for medical exemptions, and in all about a third of the youths avoided military service legally as physically unfit. For obvious reasons, married men were not exempt from the draft, but two other means of avoiding induction remained, apart from drawing a high number: the wealthy could purchase a replacement, and the poor could flee.

Initially the regime had a bad conscience about allowing replacements, since this made its rhetoric about the duties of citizenship sound hollow. But to placate prosperous peasants and notables, it did permit the hiring of a replacement under strict guidelines that made it difficult and costly but not impossible. Between 5 and 10 percent of all French draftees were replacements. For Napoleon's



Europe in 1812.

Draft evasion

prefects, the annual conscription levy was the top priority and draft evasion the number one problem in most departments. Persistence, routine stepped-up policing, and coercion gradually overcame the chronic resistance. Heavy fines levied against the parents of evaders did little good since most were too poor to pay them. But billeting troops in their homes was more effective, and, if they could not feed the troops, the community's wealthier taxpayers were required to do so. By 1811 the regime had broken the habit of draft evasion even as the demand for conscripts increased. In 1812, prefects all over France reported that the annual levies were less contentious than ever.

Napoleon had begun by drafting 60,000 Frenchmen annually, but by 1810 the quota hit 120,000, and the first of many "supplementary levies" was decreed to call up men from earlier classes who had drawn high numbers. In January 1813, after the Russian disaster, Napoleon replenished his armies by calling up the class of 1814 a year early and by repeated supplementary levies. Because he could still rely on his conscription machine, Napoleon consistently rebuffed offers by the allies to negotiate peace. Only after he lost the decisive Battle of Leipzig in October 1813 and was driven back across the Rhine did the machine break down. His call of November 1813 for 300,000 more men went largely unfilled. With the troops at his disposal the emperor fought the Battle of France skillfully, but he could not stop the allies. Shortly after Paris fell, he abdicated, on April 6, 1814, and departed for the island kingdom of Elba. France was reduced to its 1792 borders, and the Bourbons returned to the throne. Altogether—along with large levies of Italians, Germans, and other foreigners from the annexed territories and satellite states—nearly 2.5 million Frenchmen had been drafted by Napoleon, and at least 1 million of these conscripts never returned.

The most sympathetic explanation for Napoleon's relentless aggression holds that he was responding to the irreducible antagonism of Britain: French power and glory were the only antidotes to John Bull's arrogance. Others have argued that Napoleon's vendetta against Britain was merely a rationalization for a mad 10-year chase across Europe to establish a new version of Charlemagne's empire. This "imperial design" thesis, however, makes sense only in 1810, as a way Napoleon might have organized his conquests and not as the motivation for them. (Only retrospectively did Napoleon write: "There will be no repose for Europe until she is under only one Head . . . an Emperor who should distribute kingdoms among his lieutenants.") In the end one is thrown back on the explanation of temperament. In his combination of pragmatism and insatiable ambition, this world-historic figure remains an enigma. Increasingly "aristocratic" at home and "imperial" abroad, Napoleon was obviously something more than the "general of the Revolution." And yet, with Civil Code in one hand and sabre in the other, Napoleon could still be seen by Europeans as a personification on both counts of the French Revolution's explosive force.

NAPOLEON AND THE REVOLUTION

The Revolutionary legacy for Napoleon consisted above all in the abolition of the ancien régime's most archaic features—"feudalism," seigneurialism, legal privileges, and provincial liberties. No matter how aristocratic his style became, he had no use for the ineffective institutions and abuses of the ancien régime. Napoleon was "modern" in temperament as well as destructively aggressive. But in either guise he was an authoritarian with little patience for argument, who profited from the Revolution's clearing operations to construct and mobilize in his own fashion. His concept of reform exaggerated the Revolution's emphasis on uniformity and centralization. Napoleon also accepted the Revolutionary principles of civil equality and equality of opportunity, meaning the recognition of merit. Other rights and liberties did not seem essential. Unlike others before him who had tried and failed, Napoleon terminated the Revolution, but at the price of suppressing the electoral process and partisan politics altogether. Toward the end of the empire, his centralizing vision took over completely, reinforcing his personal will to power. France was merely a launching pad for Napoleon's boundless military and imperial ambition, its prime function being to raise men and money for war. In utter contrast to the Revolution, then, militarism became the defining quality of the Napoleonic regime.

The Battle of Waterloo

Napoleon's ambiguous legacy helps explain the dizzying events that shook France in 1815. When Louis XVI's brother returned as King Louis XVIII, he courted popularity by ending conscription and by agreeing to rule under a constitution (called the Charter), which provided for legislative control over budgets and taxes and guaranteed basic liberties. But the Bourbons alienated the officer corps by retiring many at half pay and frightened many citizens by not making clear how much of their property and power the church and émigrés would regain. As the anti-Napoleonic allies argued among themselves about the spoils of war, Napoleon slipped back to France for a last adventure, believing that he could reach Paris without firing a shot. At various points along the way, troops disobeyed royalist officers and rallied to the emperor, while Louis fled the country. Between March and June 1815—a period known as the Hundred Days—Napoleon again ruled France. Contrary to his expectation, however, the allies patched up their differences and were determined to rout "the usurper." At the Battle of Waterloo (June 18, 1815) British and Prussian forces defeated Napoleon's army decisively, and he abdicated again a few days later. Placed on the remote island of Saint Helena in the South Atlantic, he died in 1821. The "Napoleonic legend"—the retrospective version of events created by Napoleon during his exile—burnished his image in France for decades to come. But in the final analysis Napoleon's impact on future generations was not nearly as powerful as the legacy of the French Revolution itself. (I.Wo.)

France since 1815

THE RESTORATION AND CONSTITUTIONAL MONARCHY

Constitutionalism and reaction, 1815–30. *Louis XVIII, 1815–24.* King Louis XVIII's second return from exile was far from glorious. Neither the victorious powers nor Louis's French subjects viewed his restoration with much enthusiasm, yet there seemed to be no ready alternative to Bourbon rule. The allies avenged themselves for the Hundred Days by writing a new and more severe Treaty of Paris. France lost several frontier territories, notably the Saar basin and Savoy (Savoie), that had been annexed in 1789–92; a war indemnity of 700 million francs was imposed; and, pending full payment, eastern France was to be occupied by allied troops at French expense.

The reactionaries

Within France, political tensions were exacerbated by Napoleon's mad gamble and by the mistakes committed during the First Restoration. The problem facing the Bourbons would have been difficult enough without these tensions—namely, how to arrive at a stable compromise between those Frenchmen who saw the Revolutionary changes as irreversible and those who were determined to resurrect the ancien régime. The reactionary element, labeled ultraroyalists (or simply "ultras"), was now more intransigent than ever and set out to purge the country of all those who had betrayed the dynasty. A brief period of "white terror" in the south claimed some 300 victims; in Paris, many high officials who had rallied to Napoleon were dismissed, and a few eminent figures, notably Marshal Michel Ney, were tried and shot. The king refused, however, to scrap the Charter of 1814, in spite of ultra pressure. When a new Chamber of Deputies was elected in August 1815, the ultras scored a sweeping victory; the surprised king, who had feared a surge of antimonarchical sentiment, greeted the legislature as *la chambre introuvable* ("the incomparable chamber"). But the political honeymoon was short-lived. Louis was shrewd enough, or cautious enough, to realize that ultra policies would divide the country and might in the end destroy the dynasty. He chose as ministers, therefore, such moderate royalists as Armand-Emmanuel du Plessis, Duke de Richelieu, and Élie Decazes—men who knew the nation would not tolerate an attempt to resurrect the 18th century.

There followed a year of sharp friction between these moderate ministers and the ultra-dominated Chamber—friction and unrest that made Europe increasingly nervous about the viability of the restored monarchy. Representatives of the occupying powers began to express their concern to the king. At last, in September 1816, his ministers persuaded him to dissolve the Chamber and order new elections, and the moderate royalists emerged with a clear majority. In spite of ultra fury, several years of relative stability ensued. Richelieu and Decazes, with solid support in the Chamber, could proceed with their attempt to pursue a moderate course. By 1818 they were able, thanks to loans from English and Dutch bankers, to pay off the war indemnity and thus to end the allied occupation; at the Congress of Aix-la-Chapelle France was welcomed back into the concert of Europe. In domestic politics there were some signs that France might be moving toward a British-style parliamentary monarchy, even though the Charter had carefully avoided making the king's ministers responsible to the Chamber of Deputies. In the Chamber something anticipating a party system also began to emerge: ultras on the right, independents (or liberals) on the left, constitutionalists (or moderates) in the centre. None of these factions yet possessed the real attributes of a party—disciplined organization and doctrinal coherence. The most heterogeneous of all was the independent group—an uneasy coalition of republicans, Bonapartists, and constitutional monarchists brought together by their common hostility to the Bourbons and their common determination to preserve or restore many of the Revolutionary reforms.

The advance of the left

The era of moderate rule (1816–20) was marked by a slow but steady advance of the liberal left. Each year one-fifth of the Chamber faced reelection, and each year more independents won seats, despite the narrowly restricted suffrage. The ultras, in real or simulated panic, predicted disaster for the regime and the nation; but the king clung stubbornly to his favourite, Decazes, who by now was head of the government in all but name, and Decazes, in turn, clung to his middle way.

The uneasy balance was wrecked in February 1820 by the assassination of the king's nephew, Charles-Ferdinand, Duke de Berry. The assassin, a fanatical Bonapartist, proudly announced his purpose: to extinguish the royal line by destroying the last Bourbon still young enough to produce a male heir. In this aim he failed, for Caroline, Duchess de Berry, seven months later bore a son, whom the royalists hailed as "the miracle child." But the assassin did bring to an end the period of moderate rule and returned the ultras to power. In the wave of emotion that followed, the king dismissed Decazes and manipulated the elections in favour of the ultras, who regained control of the Chamber and dominated the new Cabinet headed by one of their leaders, Joseph, Count de Villèle.

This swing toward reaction goaded some segments of the liberal left into conspiratorial activity. A newly formed secret society called the Charbonnerie, which borrowed its name and ritual from the Italian Carbonari, laid plans for an armed insurrection, but their rising in 1822 was easily crushed. One group of conspirators—"the four sergeants of La Rochelle"—became heroic martyrs in the popular mythology of the French left. Subversion gave the government an excuse for intensified repression: the press was placed under more rigid censorship and the school system subjected to the clergy.

Meanwhile, the ultras were winning public support through a more assertive foreign policy. Spain had been in a state of quasi-civil war since 1820, when a revolt by the so-called liberal faction in the army had forced King Ferdinand VII to grant a constitution and to authorize the election of a parliament. The European powers, disturbed at the state of semianarchy in Spain, accepted a French offer to restore Ferdinand's authority by forcible intervention. In 1823 French troops crossed the Pyrenees and, despite predictions of disaster from the liberal left, easily took Madrid and reestablished the king's untrammeled power. This successful adventure strengthened the ultra politicians and discredited their critics. In the elections of 1824 the ultras increased their grip on the Chamber and won a further victory in September 1824 when the aged Louis XVIII died, leaving the throne to a new king who was the very embodiment of the ultra spirit.

Ascendancy of the ultras

Charles X, 1824–30. Charles X, the younger brother of Louis XVIII, had spent the Revolutionary years in exile and had returned embittered rather than chastened by the experience. What France needed, in his view, was a return to the unsullied principle of divine right, buttressed by the restored authority of the established church. The new king and his Cabinet—still headed by Villèle—promptly pushed through the Chamber a series of laws of sharply partisan character. The most bitterly debated of these laws was the one that indemnified the émigrés for the loss of their property during the Revolution. The cost of the operation—almost one billion francs—was borne by government bondholders, whose bonds were arbitrarily converted to a lower interest rate. A severe press law hamstrung the publishers of newspapers and pamphlets; another established the death penalty for sacrilegious acts committed in churches.

Along with these signs of reaction went a vigorous campaign to reassert the authority of the Roman Catholic church, which had been undermined by Enlightenment skepticism and by the Revolutionary upheaval. Under the Bourbons several new missionary orders and lay organizations were founded in an effort to revive the faith and to engage in good works. Catholic seminaries began to draw increasing numbers of students away from the state lycées. Charles X threw himself enthusiastically into the campaign for Catholic revival. The anticlericals of the liberal left were outraged, and even many moderates of Gallican sympathies were perturbed. Rumours spread that the king had secretly become a Jesuit and was planning to turn the country over to "the men in black."

King Charles and his ultra ministers might nevertheless have remained in solid control if they had been shrewd and sensitive men, aware of the rise of public discontent and flexible enough to appease it. Instead, they forged stubbornly ahead on the road to disaster. Villèle, though a talented administrator, lacked creative imagination and charismatic appeal. As the years passed, his leadership was increasingly challenged even within his own ultra majority. A bitter personal feud between Villèle and François-Auguste-René, Viscount de Chateaubriand, the most colourful of the ultra politicians, undermined both the ministry and the dynasty. This internal conflict contributed to Villèle's downfall; the elections of 1827 brought a sharp resurgence of liberal and moderate strength. The king patched together a disparate ministry of moderates and ultras headed by an obscure official, Jean-Baptiste, Viscount de Martignac. But Martignac lacked Charles's confidence and failed to win the support of the more moderate leftists in the Chamber. In 1829 the king brusquely dismissed him and restored the ultras to power.

The delayed consequences of this act were to be fatal to the dynasty. The king, instead of entrusting power to an able ultra such as Villèle or a popular one such as Chateaubriand, chose a personal favourite, Jules-Auguste-Armand-Marie, Prince de Polignac, a fanatical reactionary. The makeup of the Cabinet, which included several members of the most bigoted faction of "ultra-ultras," seemed to indicate the king's determination to polarize politics. That, in any case, was the immediate result. On the left the mood turned aggressively hostile; the republicans of Paris began to organize; an Orleanist faction emerged, looking to a constitutional monarchy headed by the king's cousin, Louis-Philippe, Duke d'Orléans. The liberal banker Jacques Laffitte supplied funds for a new opposition daily, *Le National,* edited by a young and vigorous team whose most notable member was Adolphe Thiers. A confrontation of some sort seemed inevitable.

The polarization of politics

Some of Polignac's ministers urged a royal coup d'état at once, before the rejuvenated opposition could grow too strong. Instead, the king procrastinated for several months, offering no clear lead or firm policy. When the Chamber met at last in March 1830, its majority promptly voted an address to the throne denouncing the ministry. The king retaliated by dissolving the Chamber and ordering new elections in July. Both Charles and Polignac hoped that pressure on the electors, plus foreign policy successes, might shape the outcome. Such a success was won at just the opportune moment: news came that Algiers had fallen to a French expeditionary force sent to punish the Bey for assorted transgressions. But even this brilliant victory could not divert the fury of the king's critics. The opposition won 274 seats, the ministry 143. Charles's alternatives were now clear: to substitute a moderate for Polignac and accept the role of constitutional monarch or to risk a royal coup d'état that would leave the Charter of 1814 in tatters. Without hesitation he chose the second path. King and ministers prepared a set of decrees that dissolved the newly elected Chamber, further restricted the already narrow suffrage, and stripped away the remaining liberty of the press. These July Ordinances, made public on the 26th, completed the polarization process and ensured that the confrontation would be violent.

The Revolution of 1830. The July Revolution was a monument to the ineptitude of Charles X and his advisers. At the outset, few of the king's critics imagined it possible to overthrow the regime; they hoped merely to get rid of Polignac. As for the king, he naively ignored the possibility of serious trouble. No steps were taken to reinforce the army garrison in Paris; no contingency plans were prepared. Instead, Charles went off to the country to hunt, leaving the capital weakly defended. During the three days known to Frenchmen as Les Trois Glorieuses (July 27–29), protest was rapidly transmuted into insurrection; barricades went up in the streets, manned by workers, students, and petty bourgeois citizens (some of them former members of the National Guard, which Charles, in pique, had disbanded in 1827). On July 29 some army units began to fraternize with the insurgents. The king, on July 30, consented at last to dismiss Polignac and to annul the July Ordinances; but the gesture came too late. Paris was in the hands of the rebels, and plans for a new regime were crystallizing rapidly.

As the insurrection developed, two rival factions had emerged. The republicans—mainly workers and students—gained control of the streets and took over the Hôtel de Ville, where on July 29 they set up a municipal commission. They looked to the venerable Marquis de Lafayette as their symbolic leader. The constitutional monarchists had their headquarters at the newspaper *Le National;* their candidate for the throne was Louis-Philippe, Duke d'Orléans. Louis-Philippe was at first reluctant to take the risk, fearing failure and renewed exile; Adolphe Thiers undertook the task of persuading him and succeeded. On July 31 Louis-Philippe made his way through a largely hostile crowd to the Hôtel de Ville and confronted the republicans. His cause was won by Lafayette, who found a constitutional monarchy safer than the risks of Jacobin rule; Lafayette appeared on the balcony with Louis-Philippe and, wrapped in a tricolour flag, embraced the duke as the crowd cheered. Two days later Charles X abdicated at last, though on condition that the throne pass to his grandson, "the miracle child." But the parliament, meeting on August 7, declared the throne vacant and on August 9 proclaimed Louis-Philippe "king of the French by the grace of God and the will of the nation."

The abdication of Charles X

The July monarchy. The renovated regime (often called the July monarchy or the bourgeois monarchy) rested on an altered political theory and a broadened social base. Divine right gave way to popular sovereignty; the social centre of gravity shifted from the landowning aristocracy to the wealthy bourgeoisie. The Charter of 1814 was retained but no longer as a royal gift to the nation; it was revised by the Chamber of Deputies and in its new form imposed on the king. Censorship was abolished; the tricolour was restored as the national flag, and the National Guard was resuscitated. Roman Catholicism was declared to be simply the religion "of the majority of Frenchmen," the voting age was lowered to 25, and the property qualification reduced to include all who paid a direct tax of 200 (formerly 300) francs. The suffrage was thus doubled, from about 90,000 to almost 200,000.

The new king seemed admirably suited to this new constitutional system. The "Citizen King" was reputed to be a liberal whose tastes and sympathies coincided with those of the upper bourgeoisie. He had spent the Revolutionary years in exile but was out of sympathy with the irreconcilable émigrés; and since his return, his house in Paris had been a gathering place for the opposition. Yet, in spite of appearances, Louis-Philippe was not prepared to accept the strictly symbolic role of a monarch who (in Thiers's phrase) "reigns but does not govern." His authority, he believed, rested on heredity and not merely on the will of the Chamber; his proper function was to participate actively in decision making and not merely to appoint ministers who would govern in his name. As time went by, he was increasingly inclined to choose ministers who shared his view of the royal power. The Orleanist system thus rested on a basic ambiguity about the real locus of authority.

The leadership of Guizot

In the Chamber two major factions emerged, known by the rather imprecise labels right-centre and left-centre. The former group, led by the historian François Guizot, shared the king's political doctrines; it saw the revised Charter of 1814 as an adequate instrument of government that needed no further change. The left-centre, whose ablest spokesman was the kingmaker Adolphe Thiers, saw 1830 as the beginning rather than the culmination of a process of change. It favoured restricting the king's active role and broadening the suffrage to include the middle strata of the bourgeoisie. These differences of viewpoint, combined with the king's tendency to intrigue, contributed to chronic political instability during the 1830s.

The decade of the 1830s was marked also by repeated challenges to the regime by its enemies on the right and the left and by a series of attempts to assassinate the king. Both the ultras (who now came to be called legitimists) and the republicans refused to forgive "the usurper" of 1830. In 1832 the Duchess de Berry, mother of "the miracle child," landed clandestinely in southern France in an effort to spark a general uprising; but the scheme collapsed, and most legitimists withdrew into sullen opposition. More serious was the agitation of the republicans, fed by rising labour discontent. In the most serious of these outbreaks (Lyon, 1831), 15,000 workers confronted the National Guard in the streets and suffered some 600 casualties before capitulating. Again in 1834 there were serious disturbances in Lyon and Paris. In 1836 it was the turn of the Bonapartist pretender to challenge the regime. Since Napoleon's death in 1821, a legend had rapidly taken shape around his name. No longer detested as a ruthless autocrat who had sacrificed a generation of young Frenchmen on the battlefield, he became transmuted into the Little Corporal who had risen to the heights by his own talents and had died a victim of British jealousy. The emperor's nephew Louis-Napoleon Bonaparte presented himself as the true heir; he crossed the frontier in 1836 and called on French troops in Strasbourg to join his cause. The venture failed ignominiously, as did also a second attempt on the Channel coast in 1840. Louis-Napoleon was condemned to prison for life but managed in 1846 to escape to England. Interspersed with these attempts at political risings were individual attacks on the king's person; the most elaborate of these plots was the one organized by a Corsican named Giuseppe Fieschi in 1835.

By 1840, however, the enemies of the regime had evidently become discouraged, and a period of remarkable stability followed. François Guizot emerged as the key figure in the ministry; he retained that role from 1840 to 1848. One of the first Protestants to attain high office in France, Guizot possessed many of the moral and intellectual qualities that marked this small but influential minority. Hard-working and intelligent, Guizot was devoted to the service of the king and to the defense of the status quo. He was convinced that the wealthy governing class was an ideal natural elite to which any Frenchman might have access through talent and effort. To those who complained at being excluded by the property qualification for voting and seeking office, Guizot's simple reply was "Enrichissez-vous!" ("Get rich!"). His government encouraged the process by granting railway and mining concessions to its bourgeois supporters and by contributing part of the development costs. High protective tariffs continued to shelter French entrepreneurs against foreign competition. The result was a modest economic boom during the 1840s, beginning the transformation of France from a largely rural into an industrial society.

Guizot's foreign policy

Guizot shared with Louis-Philippe a strong preference for a safe and sane foreign policy. The king, from the beginning of his reign, had cautiously avoided risks and confrontations and had especially sought friendly relations with Britain. In 1830, when the revolution in Paris inspired the Belgians to break away from Dutch rule, Louis-Philippe avoided the temptation of seeking to annex Belgium or of placing one of his sons on the Belgian throne. Again in 1840, when a crisis flared up in the Middle East and Thiers (then head of the government) took an aggressive stance that threatened to coalesce all of Europe against France, the king had found an excuse to replace his firebrand minister. Guizot continued this cautious line through the 1840s, with the single exception of an episode in Spain. A long contest involving rival suitors for the Spanish queen's hand finally tempted Guizot, in 1846, to try for a cheap diplomatic victory; it infuriated the British and helped to destroy the Anglo-French entente. One problem Guizot inherited from his predecessors was that of Algeria. Since 1830 the French had maintained an uneasy presence there, wavering between total withdrawal and expanded conquest. The decision to remain had been made in the mid-1830s; during the Guizot era, General Thomas-Robert Bugeaud de La Piconnerie's forces broke the back of Algerian resistance, pushed the native population back into the mountains, and began the process of colonizing the rich coastal plain.

THE SECOND REPUBLIC AND SECOND EMPIRE

The Revolution of 1848. The overthrow of the constitutional monarchy in February 1848 still seems, in retrospect, a puzzling event. The revolution has been called a result without a cause; more properly, it might be called a result out of proportion to its cause. Since 1840 the regime had settled into a kind of torpid stability; but it had provided the nation with peace abroad and relative prosperity at home. Louis-Philippe and his ministers had prided themselves on their moderation, their respect for the ideal of cautious balance embodied in the concept of *juste-milieu.* France seemed to be arriving at last at a working compromise that blended traditional ways with the reforms of the Revolutionary era.

The mood of restlessness

There were, nevertheless, persistent signs of discontent. The republicans had never forgiven Louis-Philippe for "confiscating" their revolution in 1830. The urban workers, moved by their misery and by the powerful social myths engendered by the Great Revolution, remained unreconciled. For a decade or more they had been increasingly drawn toward socialism in its various utopian forms. An unprecedented flowering of socialist thought marked the years 1840–48 in France: this was the generation of Barthélemy-Prosper Enfantin, Charles Fourier, Auguste Blanqui, Louis Blanc, Pierre-Joseph Proudhon, Étienne Cabet, and many others. Most of these system builders preached persuasion rather than violence, but they stimulated the hopes of the common man for an imminent transformation of society. Within the bourgeoisie as well, there was strong and vocal pressure for change in the form of a broadening of the political elite. Bills to extend the suffrage (and the right to hold office) to the middle bourgeoisie were repeatedly introduced in parliament but were stubbornly opposed by Guizot. Even the National Guard, that honour society of the lesser bourgeoisie, became infected with this mood of dissatisfaction.

Other factors, too, contributed to this mood. In 1846 a crop failure quickly developed into a full-scale economic crisis: food became scarce and expensive; many businesses went bankrupt; unemployment rose. Within the governing elite itself there were signs of a moral crisis: scandals that implicated some high officials of the regime and growing dissension among the Notables. Along with this went a serious alienation of many intellectuals. Novelists such as Victor Hugo, George Sand, and Eugène Sue glorified the common man; the caricaturist Honoré Daumier exposed the foibles of the nation's leaders; and historians such as Jules Michelet and Alphonse de Lamartine wrote with romantic passion about the heroic episodes of the Great Revolution.

Beginning in 1847, the leaders of the opposition set out to take advantage of this restless mood and to force the regime to grant liberal reforms. Since public political meetings were illegal, they undertook a series of political "banquets" to mobilize the forces of discontent. This campaign was to be climaxed by a mammoth banquet in Paris on Feb. 22, 1848. But the government, fearing violence, ordered the affair canceled. On the 22nd, crowds of protesting students and workers gathered in the streets and began to clash with the police. The king and Guizot expected no serious trouble: the weather was bad, and a large army garrison was available in case of need. But the disorders continued to spread, and the loyalty of the National Guard began to seem dubious. Toward the end of two days of rioting, Louis-Philippe faced a painful choice: to unleash the army (which would mean a bloodbath) or to appease the demonstrators. Reluctantly, he chose the second course and announced that he would replace the hated Guizot as his chief minister. But the concession came too late. That evening, an army unit guarding Guizot's official residence clashed with a mob of demonstrators, some 40 of whom died in the fusillade. By the morning of February 24, the angry crowd was threatening the royal palace. Louis-Philippe, confronted by the prospect of civil war, hesitated and then retreated once more; he announced his abdication in favour of his nine-year-old grandson and fled to England.

Abdication of Louis-Philippe

The Second Republic, 1848–52. The succession to the throne was not to be decided so easily, however. The Chamber of Deputies, invaded by a mob that demanded a republic, set up a provisional government whose members ranged from constitutional monarchists to one radical deputy, Alexandre-Auguste Ledru-Rollin. Led by the poet-deputy Alphonse de Lamartine, the members of the government proceeded to the Hôtel de Ville, where the radical republican leaders had begun to organize their own regime. After considerable palaver, the provisional government co-opted four of the radical leaders, including the socialist theoretician Louis Blanc and a workingman who called himself Albert. Under heavy pressure from the crowd surrounding the Hôtel de Ville, the government proclaimed the republic. During the next few days, continuing pressure from the social reformers pushed the government farther than its bourgeois members really wanted to go. The government issued a right-to-work declaration, obligating the state to provide jobs for all citizens. To meet the immediate need, an emergency-relief agency called the *ateliers nationaux* (national workshops) was established. A kind of economic and social council called the Luxembourg Commission was created to study programs of social reform; Louis Blanc was named its president. The principle of universal manhood suffrage was proclaimed—an almost unprecedented experiment in that day and one that increased the electorate at a stroke from 200,000 to 9,000,000. In matters of foreign policy, on the other hand, Foreign Minister Lamartine resisted radical demands. The radicals were eager for an ideological crusade on behalf of all peoples who were thirsting for freedom: Poles, Italians, Hungarians, and Germans had launched their own revolutions and needed help. Lamartine preferred to confine himself to lip-service support, since he was aware that an armed crusade would quickly inspire an anti-French coalition of the major powers.

The election of April 23, 1848

By April 23, when Frenchmen went to the polls to elect their constituent assembly, the initial mood of brotherhood and goodwill had been largely dissipated. Paris had become a caldron of political activism; dozens of clubs and scores of newspapers had sprung up after the revolution. Severe tension developed between moderates and radicals both within and outside the government and led to a number of violent street demonstrations that were controlled with difficulty. The *ateliers nationaux* satisfied no one: for the radicals they were a mere caricature of social reform, whereas for the moderates they were a wasteful and dangerous experiment that attracted thousands of unemployed to Paris from every corner of France. Financial problems plagued the government, which sought a solution by imposing a special 45-centime surtax on each franc of direct property taxes; this burden weighed most heavily on the peasantry and was bitterly resented in the countryside. The radicals, fearing that universal suffrage under these conditions might produce unpleasant results, vainly urged postponement of the elections until the new voters could be "educated" as to the virtues of a social republic.

The election returns confirmed the radicals' fears: the country voted massively for moderate or conservative candidates. Radicals or socialists won only about 80 of the 880 seats; the rest were bourgeois republicans (500) or constitutional monarchists (300). Lamartine led the popularity parade, being elected in 10 districts. When the assembly convened in May, the new majority showed little patience or caution; it was determined to cut costs and end risky experiments. In spite of Lamartine's efforts to maintain broad republican unity and avert a sharp turn to the right, the assembly abolished the Luxembourg Commission and the *ateliers nationaux* and refused to substitute a more useful program of public works to provide for the unemployed.

The June Days

The immediate consequence was a brief and bloody civil war in Paris—the so-called June Days (June 23–26, 1848). Thousands of workers suddenly cut off the state payroll were joined by sympathizers—students, artisans, employed workers—in a spontaneous protest movement. Barricades went up in many working-class sections. The assembly turned to General Louis Eugène Cavaignac as a saviour. Cavaignac had made his mark in repressing Algerian rebel tribes and was entrusted with full powers to do the same in Paris. He gave the workers time to dig themselves in, then brought up artillery against their barricades. At least 1,500 rebels were killed; 12,000 were arrested, and many were subsequently exiled to Algeria. The radical movement was decapitated; the workers withdrew into silent and bitter opposition.

Social conflict now gave way to political maneuvering and constitution making. Cavaignac was retained in office as temporary executive, while the assembly turned to its central task. After six months of discussion, it produced a constitution that appeared to be the most democratic in Europe. The president of the republic would be chosen for a four-year term by universal male suffrage; a one-house legislative assembly would be elected for three years by the same suffrage. What remained unclear was the relationship between president and assembly and the way out of a potential deadlock between them.

This problem might not have been fatal if the right kind of president had been available in 1848. Instead, the voters chose Louis-Napoleon Bonaparte, who had returned from British exile in September, after having successfully stood for the constituent assembly in a by-election. He had made a poor initial impression; indeed, some politicians, such as Thiers, backed him for the presidency because they thought him too stupid to rule and thus soon to be shunted aside for an Orleanist monarch. What he possessed, however, was a name—a name that Frenchmen knew and that conveyed an aura of glory, power, and public order. In December Louis-Napoleon won by a landslide, polling 5.5 million votes as against 2 million for all other candidates combined. In May 1849 the election of the legislative assembly produced an equal surprise. The two extremes—the radical left and the monarchist right—made impressive gains, whereas the moderate republicans, who had shaped the new system, were almost wiped out. The moderates emerged with only 80 seats, the radicals with 200, the monarchists with almost 500. But the monarchist majority lacked coherence, being split into legitimist and Orleanist factions that distrusted each other and differed on political principles.

During the next two years, President Bonaparte played his cards carefully, avoiding conflict with the monarchist assembly. He pleased Roman Catholics by restoring the pope to his temporal throne in Rome, from which he had been driven by Roman republicans. At home, he accepted without protest a series of conservative measures adopted by the assembly: these laws deprived one-third of all Frenchmen of the right to vote, restricted the press and public assemblage, and gave the church a firm grip on public as well as private education. Yet there was some reason to doubt that Louis-Napoleon really welcomed this trend toward conservatism. His writings of the 1840s had been marked by a kind of technocratic outlook, in the tradition of Saint-Simonian socialism. His effort to please the assembly probably derived from his hope that the assembly would reciprocate: he wanted funds from the treasury to pay his personal debts and run his household, along with a constitutional amendment that would allow him to run for a second term.

Louis-Napoleon's coup d'état

By 1851 it was clear that the majority was not ready to give the president what he wanted. His alternatives were to step down in 1852, bereft of income and power, or to prepare a coup d'état. Some members of his entourage had long urged the latter course; Louis-Napoleon now concurred, with some reluctance.

On the early morning of Dec. 2, 1851, some 70 leading politicians were arrested, and the outlines of a new constitution were proclaimed to the nation. It restored manhood suffrage, sharply reduced the assembly's powers, and extended the president's term to 10 years. Although the coup went off smoothly, it was followed by several days of agitation. Barricades went up in the streets, crowds clashed with troops and police in Paris and in the provinces, several hundred demonstrators were killed, and 27,000 were arrested. Once the resistance was broken, Louis-Napoleon proceeded with his announced plebiscite on the new constitution and was gratified to receive the approval of 92 percent of those who voted. But the authoritarian republic was only a stopgap. Officially inspired petitions for a restoration of the empire began to flow to Paris; the Senate responded to what it described as the nation's desires, and on Dec. 2, 1852, Napoleon III was proclaimed emperor of the French. This time there was no open protest; and the voters, in a new plebiscite, accorded Napoleon a handsome majority of 97 percent.

The Second Empire, 1852–70. Posterity's image of Napoleon III and his regime has not been uniform. Some historians have seen him as a shallow opportunist whose only asset was a glorious name. Others have described him as a visionary reformer and patron of progress, a man who successfully attempted to reconcile liberty and authority, national prestige and European cooperation. The emperor's enigmatic character and the contradictions built into his regime make it possible to argue either case.

The authoritarian years. From 1852 to 1859 the empire was authoritarian in tone. Civil liberties were narrowly circumscribed; vocal opponents of the regime remained in exile or were constrained to silence; parliament's wings were clipped; elections to the Legislative Corps (Corps Législatif; the lower house of the Parliament) were spaced at six-year intervals and were "managed" by Napoleon's prefects, who sponsored official candidates. An illusion of popular control was created by the use of the plebiscite to ratify decisions already made. The emperor and his ministers (members of his personal entourage or former Orleanist politicians) rested their authority on the peasant masses, the business class, the church, and those local notables who were willing to cooperate. Little attempt was made to install a new power elite or to create an organized Bonapartist party. Policy during the 1850s was consistently conservative; defense of the social order took precedence over reform.

Economic growth during the Second Empire

The most striking achievements of these authoritarian years were in economic growth and foreign policy. France had never before experienced such vigorous economic expansion. During the Second Empire industrial production doubled, foreign trade tripled, the use of steam power increased fivefold, railway mileage sixfold. The first great investment banks were founded (*e.g.*, the Péreire brothers' Crédit Mobilier) and the first department store (the Bon Marché in Paris). The surge of French enterprise transcended frontiers: French capital and engineers built bridges, railways, docks, and sewerage systems throughout much of continental Europe.

In part, this burst of energy had its source in favourable world conditions: the availability of more rapid steam transportation, an influx of new gold from overseas, general recovery from the slump of 1846–51. But to some degree Napoleon's government could claim credit, too—not so much by direct intervention in economic life as by creating a favourable climate for private enterprise. Many Frenchmen took advantage of the opportunities offered; they accumulated sizable fortunes and founded enterprises that still exist today. Among these entrepreneurs, however, there was a disproportionate number of "outsiders"—notably men of Protestant or Jewish origin, or former disciples of Henri de Saint-Simon. Alongside these dynamic newcomers, the older business and banking leaders continued to operate on more cautious traditional lines. From the Second Empire onward, the French economy would combine these two contrasting sectors: a dynamic modernized element superimposed upon a largely static traditional kind of enterprise.

Napoleon's foreign policy at the outset was cautious; "the empire means peace," he assured his countrymen and the nervous powers of Europe. Yet for a ruler who bore the name Napoleon the prudent and colourless policy of a Louis-Philippe seemed hardly appropriate. Besides, the emperor was eager to achieve recognition from the other European monarchs, who regarded him as an upstart. It was for these reasons rather than because of urgent national interest that he becamc involvcd in the Crimean War in 1854. Britain and Russia were engaged in a contest for influence in the crumbling Turkish empire. A dispute over the holy places in Palestine gavc Napoleon an excuse to offer the British his support and thus to restore the Franco-British ententc. Although the Crimean campaign was on the whole a fiasco for all of thc participating armies, the French forces came off less ingloriously than the others and could with some justicc pose as victors. Napoleon served as host for thc Paris peace conference that ended the war in 1856. Midway through thc conference, the birth of a male heir to the emperor and his empress, Eugćnie, seemed to assure the pcrmanence of the dynasty.

The liberal years. The empire thus appeared to have compiled a record of unbroken succcsses and to be beyond challenge by its domestic critics. Perhaps it was this stability and self-confidence that led Napoleon, beginning in 1859, to turn in the dirction of libcralizing the empire. The immediate impulse for this dramatic reversal was the attempted assassination of the emperor in January 1858 by an Italian patriot, Felice Orsini, who sought thus to draw public attention to the frustrated hopes of Italian nationalists. Napoleon, shaken by the cpisode and by the reminder that in his youth he, too, had fought for Italian independence, met secretly in July 1858 with Count Cavour, premier of Piedmont; thc two mcn laid plans designed to evict Austria from northern Italy and to convert Italy into a confederation of states headed by the pope. In return, France was promised Nicc and Savoy (Savoie). The new allies provoked the Austrians into a declaration of war in April 1859, and Napoleon led his armies across the Alps. French victories at Magenta and Solferino were followed by a somewhat premature scttlement in which the Austrians turned over the province of Lombardy to the Piedmontese. The campaign had aroused the passions of Italian nationalists all up and down the peninsula; revolutions broke out in some of the smaller Italian states, and in 1860 the colourful freebooter Giuseppe Garibaldi set forth from Piedmont to conquer Sicily and Naples.

Intervention in Italy

These repercussions of Napoleon's new foreign policy stirred up bitter controversy in France. Conservatives were outraged and feared that the pope would be deposed as temporal ruler of Rome by the Italian nationalists. On the other hand, the long-silent liberal and radical opposition voiced reluctant approval. It is likely that Napoleon, whose bent toward Saint-Simonian reform ideas was strong, had never been very comfortable in his alliance with the conservatives and welcomed a chance to indulge his deeper instincts. At any rate, late in 1859 he announced the first hesitant steps toward a liberal empire. Political exiles were amnestied, press controls were relaxed, and the Legislative Corps was given slightly increased authority. An even more dramatic turn toward economic liberalism soon followed; in January 1860 Napoleon negotiated a low-tariff treaty with Britain, ending the long tradition of protectionism that had insulated French producers. By this move, however, the emperor alienated the businessmen, who until now had been his strong supporters.

Some of the emperor's advisers had sharply opposed the turn toward liberalism. Events during the next decade seemed to confirm their warnings; for the empire now ran into increasingly stormy weather. The political opposition, stifled since 1851, showed little gratitude to its benefactor and took every opportunity to harass the government. In the 1863 elections, opposition candidates polled two million votes, and 35 of them were elected to the Legislative Corps—including such effective spokesmen as the Orleanist Thiers and the republican Jules Favre. A downward turn in the economy played into the hands of the opposition. Foreign policy errors added to the regime's embarrassment: Napoleon's ill-conceived intervention in Mexico, where he hoped to establish a client empire under Maximilian of Austria, proved costly and futile and seemed to threaten a conflict with the United States. And from the mid-1860s a new threat began to loom up across the Rhine: the burgeoning power of Prussia, under the guidance of Otto von Bismarck.

Further liberal reforms

Despite these evil portents, Napoleon clung doggedly to his liberalization venture; additional reforms were granted throughout the decade. He expressed sympathy with the workers, granted them a kind of extralegal right to form trade unions and to strike, and helped them organize mutual-aid societies. His minister of education, Victor Duruy, carried out an enlightened program of broadened public education, including the establishment of the first secondary education for girls. In 1867 the emperor restored quite considerable freedom of the press and of public assembly and further broadened the powers of the Legislative Corps. Yet the response of the voters to these concessions caused some dismay; in the elections of 1869 the opposition vote rose to 3.3 million, and the number of seats held by oppositionists more than doubled.

The emperor now faced a momentous choice: a still further dose of liberalism or a brusque return to the authoritarian empire. He chose the former alternative; in January 1870 he asked the leader of the liberal opposition, Émile Ollivier, to form a government. Ollivier supervised the drafting of a new constitution, which, though hybrid in nature, converted the empire into a quasi-parliamentary regime. The ministers were declared to be "responsible," and their powers (as well as those of the Legislative Corps) were increased. At the same time the emperor retained most of his existing prerogatives, so that the real locus of power in case of a conflict was unclear. Nevertheless, the voters, when consulted by referendum (May 8, 1870), gave the new system a massive vote of confidence: 7 million in favour to only 1.5 million against. Outwardly, at least, it appeared that the emperor had found a widely accepted solution. But war and defeat only four months later were to prevent a fair test of the liberal empire in its final form.

The Franco-German War. Napoleon, meanwhile, had become uncomfortably involved in a diplomatic poker game with Bismarck. Prussian victories over Denmark (1864) and Austria (1866) indicated a serious shift in the European balance of power. Napoleon, aware that he faced a severe challenge, set out to strengthen his armed forces; he proposed a tighter conscription law that would increase the size of the standing army but had to retreat in the face of public and parliamentary hostility. The crisis that finally erupted in July 1870 over the succession to the Spanish throne was clumsily handled by French officials. The French successfully blocked the accession of a Hohenzollern prince in Spain, then demanded further guarantees for the future; they thus provided Bismarck with an easy opportunity to arouse German opinion and to goad France into declaring war on July 19.

Few French or foreign observers anticipated the military disaster that followed. The French armies, sunk in routine and slow to mobilize, were not yet ready to fight when the Prussian forces under Helmuth von Moltke crossed into France. One French army, under Achille-François Bazaine, was bottled up in Metz; another, under Patrice Mac-Mahon, was cornered at Sedan. There, on September 1, the Prussians won a clear-cut victory; Napoleon himself was taken prisoner. The regime could not survive such a humiliation. When the news reached Paris on September 4, crowds filled the streets and converged on the Legislative Corps, demanding the proclamation of a republic. The imperial officials put up no serious resistance; the Revolution of September 4 was the most bloodless in French history.

The Revolution of September 4

THE THIRD REPUBLIC

A provisional government of national defense was set up and took as its first task the continuation of the war against the invaders. Composed of the deputies representing Paris and formally headed by General Louis-Jules Trochu, the new government's most forceful member was Léon Gambetta, hero of the radical republicans. Gambetta, a young Parisian lawyer of provincial origin, had been elected to the Legislative Corps in 1869 and had already made his mark through his energy and eloquence. As minister of the interior and, some weeks later, minister of war as well, he threw himself into the task of improvising military resistance. His task was complicated by the advance of the Prussian forces, which, by September 23, surrounded and besieged Paris. Gambetta shortly left the city by balloon to join several members of the government at Tours. During the next four months, Gambetta's makeshift armies fought a series of indecisive battles with the Prussians in the Loire valley and eastern France. But his attempt to send a force northward to relieve Paris from siege was frustrated by Moltke. Adolphe Thiers had been sent meanwhile to tour the capitals of Europe in search of support from the powers; but he returned empty-handed. By January 1871 it was clear that further armed resistance would be futile. Over Gambetta's angry protests, an armistice was signed with the Prussians on January 28.

One provision of the armistice called for the prompt election of a national assembly with authority to negotiate a definitive treaty of peace. That election, held on February 8, produced an assembly dominated by monarchists—more than 400 of them, compared to only 200 republicans and a few Bonapartists. The decisive issue for the voters, however, had not been the nature of the future regime but simply war or peace. Most of the monarchists had campaigned for peace; the republicans had insisted on a last-ditch fight. Most Frenchmen opted for peace, though Paris and certain provinces, such as Alsace, voted heavily for republicans. When the National Assembly convened in Bordeaux on February 13, it chose the aging Orleanist Adolphe Thiers as "chief of the executive power of the French republic." Thiers had been the most outspoken critic of Napoleon's foreign policy and had repeatedly warned the country of the Prussian danger. He set out at once to negotiate a settlement with Bismarck; on March 1 the Treaty of Frankfurt was ratified by a large majority of the assembly. The terms were severe: France was charged a war indemnity of five billion francs plus the cost of maintaining a German occupation army in eastern France until the indemnity was paid. Alsace and half of Lorraine were annexed to the new German Empire. The German army was authorized to stage a victory march through the Arc de Triomphe de l'Étoile in Paris. After the assembly ratified the treaty, the deputies of the lost provinces (including Léon Gambetta) resigned their seats in protest.

The Treaty of Frankfurt

The Paris Commune. A few days later, the assembly transferred the seat of government from Bordeaux to Versailles. It had scarcely arrived when it was confronted by a major civil war—the rebellion of the Paris Commune. This event, complex in itself, has been made even more difficult to understand by the mythology that later grew up around it. Karl Marx, who promptly hailed the Commune as the first great uprising of the proletariat against its bourgeois oppressors, was partly responsible for creating this mythology. There was undoubtedly a class-struggle element in the episode, but this was not the central thread. Parisians, tense and irritable after the long strain of the siege, were outraged by the action of rural France in electing a monarchist assembly committed to what they regarded as a dishonourable peace. They were further angered by the assembly's subsequent acts, notably those that ended the wartime moratorium on debts and rents, cut off further wage payments to the National Guard (which had been resuscitated in Paris after the empire fell), and transferred the capital to Versailles rather than to Paris.

Thiers, aware that Paris was in an ugly mood, thought it prudent to disarm the National Guard, which heavily outnumbered the regular army units at the government's disposition. Before dawn on March 18 he sent troops to confiscate the National Guard cannon on the butte of Montmartre. A crowd gathered; a bloody encounter ensued; two generals were caught and lynched by the mob. As violence spread through the city, Thiers hastily withdrew all troops and government offices from Paris and went to Versailles to plan his strategy. He appealed successfully to Bismarck to release French prisoners of war in order to form a siege army that could eventually force Paris to capitulate. During the next two months, this governmental force was slowly assembled. Within Paris, meanwhile, initial chaos gradually gave way to an improvised experiment in municipal self-government. On March 26, Parisians elected a council that promptly adopted the traditional label Commune of Paris. Its membership ranged from radical republicans of the Jacobin and Blanquist variety to socialists of several different sorts—notably disciples of Proudhon, who favoured a decentralized federation of self-governing communes throughout France. These internal divisions prevented any vigorous or coherent experiments in social reform and also interfered with the Commune's efforts to organize an effective armed force. Communes on the Paris model were set up briefly in several other cities (Lyon, Marseille, Toulouse) but were quickly suppressed.

"Bloody Week"

By May 21 Thiers's forces were ready to strike. In the course of "Bloody Week" (May 21–28) the Communards resisted, street by street, but were pushed back steadily to the heart of Paris. In their desperation, they executed a number of hostages (including the archbishop of Paris) and in the last days set fire to many public buildings, including the Tuileries Palace and the Hôtel de Ville. A final stand was made in Père-Lachaise Cemetery, where the last resisters were shot down against the Mur des Fédérés—ever since, a place of pilgrimage for the French left. Thiers's government took a terrible vengeance. Twenty thousand Communards were killed in the fighting or executed on the spot; thousands of survivors were deported to the penal islands, while others escaped into exile.

The formative years (1871–1905). The repression of the Paris Commune left its mark on the emerging republic. The various socialist factions and the newly organized labour movement were left leaderless; the resultant vacuum eventually opened the way to Marxist activists in the 1880s. Much of the working class became more deeply alienated than before, but, to the moderate and conservative elements, Thiers gained added stature as the preserver of law and order against "the reds." His ruthless action probably hastened the conversion of many rural and small-town Frenchmen to the idea of a republic because the regime had proved its toughness in handling subversion. A large number of by-elections to the assembly in July 1871 brought startling gains to the republicans: they won 99 of 114 vacancies. The voters were clearly willing to accept a republic so long as it was run by a man like Thiers.

Republican gains

Attempts at a restoration. The monarchists, however, still held a comfortable majority in the assembly and continued to hope and plan for a restoration. Legitimists and Orleanists remained at odds, but a compromise seemed possible. The Bourbon pretender, the Count de Chambord ("the miracle child" of 1820), was old and childless; the Orleanist pretender, Philippe, Count de Paris, was young and prolific. The natural solution was to restore Chambord, with the Count de Paris as his successor. Chambord, however, refused to accept the throne except on his own terms, which implied a return to the principle of absolute royal authority, unchecked by constitutional limitations. The Orleanists and even some legitimists found this too much to swallow. For the time being, they, too, settled for Thiers's presidential rule.

During the next two years, Thiers's position was beyond challenge, and he gave the republic vigorous and efficient leadership. He reorganized the army and worked to restore national morale; he successfully floated two bond issues that permitted the war indemnity to be paid off in 1873, thus ending the German occupation ahead of schedule. Late in 1872, however, Thiers abjured his long-held Orleanist faith and publicly announced his conversion to republicanism. The monarchists, outraged and seeing their majority in the assembly dwindling because of by-elections, found an excuse to force Thiers's resignation as provisional president (May 1873) and hastily substituted the commander of the army, Marshal Patrice de Mac-Mahon. Behind the scenes, monarchist politicians again set out to arrange an agreement between the two pretenders. Their hopes were once more sabotaged by Chambord, who again announced that he would return only on his own terms and under the fleur-de-lis flag of the old regime. The disheartened monarchists fell back on waiting for the Bourbon line to die out. But when Chambord passed from the scene in 1883, it was too late for a restoration.

The constitution of the Third Republic. Meanwhile, the task of writing a constitution for the republic could no longer be postponed. The assembly began its deliberations in 1873; in 1875 it adopted a series of fundamental laws, which, taken collectively, came to be known as the constitution of the Third Republic. A patchwork compromise, it established a two-house legislature (with an indirectly elected Senate as a conservative check on the Chamber of Deputies); a Council of Ministers (Cabinet), responsible to the Chamber; and a president, elected for seven years by the two houses, with powers resembling those of a constitutional monarch. The label republic was approved by a single-vote margin. Monarchists believed that this system could be easily converted to their purposes once the right monarch was available. The constitution left untouched many aspects of the French governmental structure, notably the centralized administrative system inherited from Napoleon I, the hierarchy of courts and judges, and the Concordat of 1801, governing church-state relations. At the end of 1875 the National Assembly at last dissolved itself, and the provisional phase of the Third Republic came to an end.

End of the provisional phase

The new Senate, which heavily overrepresented rural France, was safely monarchist from the outset; and the term of President Mac-Mahon, a loyal monarchist, ran until 1880. But when the first Chamber of Deputies was elected in 1876, the republicans won more than two-thirds of the seats. A period of severe friction between Mac-Mahon and the Chamber followed, and a crisis in May 1877 produced a total deadlock. Mac-Mahon dissolved the Chamber and called on the voters' support, but again they opted for the republic, by a narrower but clear-cut margin. Léon Gambetta, who had returned to political life and had led the republicans during the campaign, called on Mac-Mahon to "give in or get out." The president gave in, naming a premier acceptable to the republican majority. Two years later partial elections gave the republicans control of the Senate, and Mac-Mahon shortly found an excuse to resign. He was replaced by a colourless republican, Jules Grévy, who was believed to favour a reduced role for the president.

Republican factions. With the republican regime apparently safe from outside attack, rival factions developed among the republicans. During the 1880s the labels Radical and Opportunist began to be attached to the two wings of the republican movement. On the left, the Radicals saw themselves as heirs to the Jacobin tradition: they stood for a strong centralized regime, intransigent anticlericalism, an assertive nationalism in foreign policy, a revision of the constitution to prune out its monarchical aspects, and such social reforms as labour laws and a graduated income tax; their most colourful spokesman was Georges Clemenceau, a ferocious debater and duelist who specialized in overthrowing Cabinets. The Opportunists (so named by a satirical journalist because of their penchant for compromises and postponements) occupied the centre seats in the Chamber: their stance was more cautious and their techniques gradualist; they were content to work within the system, and they aimed to restrict governmental interference in the affairs of private citizens. Only on the issue of the church's role in politics and education were the two factions in general agreement.

Opportunist control. Between 1879 and 1899 the Opportunists, with only brief interruptions, controlled the machinery of government. Gambetta, their most dynamic leader, had begun his career as an outspoken Radical, but in time his political instincts had prevailed. The other Opportunist leaders—men such as President Grévy and Jules Ferry—disliked Gambetta's flamboyance, however, and feared his alleged dictatorial ambitions; they kept him out of the premiership save for a brief interlude in 1881–82, shortly before his death. Ferry served as premier or in other key Cabinet posts during most of the period from 1880 to 1885 and left his mark on two institutions: the public school system and the colonial empire. His school laws made primary education free, compulsory, and secular, with religious teaching in the public schools replaced by "civic education"; a strong anticlerical bias thenceforth marked French public education. Ferry's support of various colonial expeditions—sometimes behind the back of the Chamber—gave France protectorates over Tunisia and Annam and Tongking (Tonkin), a large new colony in the Congo basin, and an initial foothold in Madagascar. This expansionist policy, unpopular at the time, led later generations to call Ferry the founder of the French empire.

Educational reforms and colonial expansion

In the 1885 elections the monarchists, Bonapartists, and Radicals all made significant gains, partly because of boredom with the Opportunists, Catholic resentment over the school laws, and revived agitation by socialist organizers. The Opportunists, lacking a clear majority in the Chamber, sought Radical support to form a Cabinet; the Radicals insisted on the inclusion of General Georges Boulanger as minister of war. Within a few weeks Boulanger was the most talked-about man in France. He restored the tradition of military parades and rode at their head; he instituted popular reforms in the army; and he spoke out in chauvinistic fashion against the Germans, thus reviving the memory of 1871 and the lost provinces. The unnerved Opportunist leadership dropped him from the Cabinet and sent him in 1887 to an obscure provincial command. But Boulanger's backers urged him to plunge into politics and began to enter his name in by-elections. Privately, monarchist and Bonapartist agents also made contact with Boulanger, promising financial support and hoping to use him for their cause.

By 1889 the Boulanger movement had become a major threat to the regime. The government had placed him on the retired list, but this merely freed him to run openly for office on a vague program of constitutional revision. He triumphed in a series of by-elections, but his goal was the parliamentary election of 1889, which he hoped to turn into a kind of national plebiscite. Just prior to the election, however, believing that he was about to be arrested for subversive activities, Boulanger took flight to Brussels. His movement gradually disintegrated; word leaked out of his dealings with the monarchists, and his supporters fell away. The Opportunists' hold on the republic was strengthened by the discomfiture of those on both right and left who had been taken in by this adventurer.

The Panama Scandal

A new crisis soon arose for the regime: the Panama Scandal. Ferdinand de Lesseps, the noted French engineer who had built the Suez Canal, had organized a joint-stock company to cut a canal across the Isthmus of Panama. The venture proved difficult and costly; in 1889 the company collapsed, and large numbers of shareholders were stripped of their savings. Demands for a parliamentary investigation proved ineffective until 1892, when a muckraking journalist named Édouard Drumont obtained evidence that agents of the company had bribed a large number of politicians and journalists in a desperate effort to get funds to keep the company afloat. The directors of the company and several deputies and senators were brought to trial in 1893, but the outcome was on the whole a whitewash. The regime survived the scandal, but the effects were more serious than appeared to be the case. Cynicism about the honesty of the republic's political leadership gave added strength to the rising socialist movement; in 1893 almost 50 socialists won seats in the Chamber. Georges Clemenceau, unjustly accused of involvement in the scandal, was defeated; and many prominent Opportunists, tainted by the affair, withdrew and were replaced by such younger men as Raymond Poincaré and Louis Barthou, who thenceforth preferred to call themselves Progressists or Moderates.

The dramatic socialist gains in 1893 resulted only partly from the Panama Scandal. For more than a decade socialism had been gaining strength among the increasingly class-conscious urban workers. The movement was weakened, however, by multiple splits into antagonistic factions. The Marxist party created by Jules Guesde in 1880 broke up two years later into Guesdists and followers of Paul Brousse—the latter group popularly called possibilists because of their gradualist temper. In 1890 a third faction broke away, headed by Jean Allemane and limited to simon-pure proletarian members. Alongside these Marxist sects there were the Blanquistes (disciples of Auguste Blanqui), the anarchists (whose terrorist campaign in the early 1890s earned them wide notoriety), and a considerable scattering of independent socialists (mainly intellectuals, notably Jean Jaurès). By 1900 the parties had been reduced to the two led by Guesde and Jaurès, which merged in 1905 to form the French Section of the Workers' International (Section Française de l'Internationale Ouvrière [SFIO]), known as the Socialist Party.

The trade union movement, however, refused to join forces with the socialists. Trade unions were finally legalized in 1884 and joined together to form a national General Labour Confederation (Confédération Générale du Travail [CGT]) in 1895. CGT leaders rejected political action in favour of direct action—sabotage, boycotts, strikes, and especially the general strike, which they saw as the ultimate weapon that would transform France into a workers' state. This doctrine, known as revolutionary syndicalism, made the French trade union movement appear to be one of the most radical in Europe. In practice, however, the trade union rank and file was less revolutionary than its leadership.

Revolutionary syndicalism

The Dreyfus affair. The 1890s also saw the Third Republic's greatest political and moral crisis—the Dreyfus affair. In 1894 Captain Alfred Dreyfus, a career army officer of Jewish origin, was charged with selling military secrets to the Germans. He was tried and convicted by a court-martial and sentenced to life imprisonment on Devil's Island off the South American coast. Efforts by the Dreyfus family to reopen the case were frustrated by the general belief that justice had been done. But secrets continued to leak to the German embassy in Paris, and a second officer, Major Marie-Charles-Ferdinand Esterhazy, became suspect. The chief of army counterintelligence, Colonel Georges Picquart, eventually concluded that Esterhazy and not Dreyfus had been guilty of the original offense, but his superior officers refused to reopen the case. Rumours and scraps of evidence soon began to appear in the press; and a few politicians, notably Georges Clemenceau, took up Dreyfus' cause. But the army high command refused to discuss the affair, although army officers leaked documents to the press in an effort to discredit the critics. Each leak aroused new controversy, and by 1898 the case had become a violently divisive issue. Intellectuals of the left led the fight for Dreyfus, while right-wing politicians and many Roman Catholic periodicals defended the honour of the army. The socialists were split: Jaurès insisted that no socialist could remain aloof on such a moral issue, while Guesde called the conflict a bourgeois squabble.

In 1898 some of the army's most persuasive documents against Dreyfus were discovered to be forgeries. Esterhazy promptly fled to England. In a second court-martial, late in 1899, Dreyfus was again found guilty but with extenuating circumstances; he received a presidential pardon and was later (1906) vindicated by a civilian court. For a generation the affair left deep scars on French political and intellectual life. The Moderates, who had tried to avoid involvement in the affair and in the end had split into two warring factions, lost control to the Radicals.

Radical control

A coalition Cabinet headed by René Waldeck-Rousseau, a pro-Dreyfus Moderate, took office in June 1899; the Radicals dominated the coalition, and even the socialists supported it. From then until the end of the Third Republic the Radical Party (thenceforth called Radical-Socialist) remained the fulcrum of French political life. Both the army and the church were seriously hurt by their role in the affair; republicans of the left were more convinced than ever that both institutions were antirepublican and hostile to the rights of man enunciated during the Revolution. The new left majority retaliated by bringing the army under more rigorous civilian control and by embarking on a new wave of anticlerical legislation. Most religious orders were dissolved and exiled, and in 1905 a new law separated church and state, thus liquidating the Concordat of 1801.

Foreign policy. Meanwhile, some important successes were being scored in the field of foreign policy. For two decades after 1871 France had remained diplomatically isolated in Europe. Bismarck, to ward off potential ideas of revenge in France, had shrewdly encouraged French governments to embark on colonial conquest overseas and had negotiated alliances with all those European powers the French might otherwise have courted. He thus kept Austria-Hungary, Russia, and Italy in tow, while Britain chose to remain aloof in "splendid isolation." Upon Bismarck's fall in 1890, the German emperor William II terminated the secret treaty between Germany and Russia. The Russians began to cast about for friends and looked with some distaste toward Paris. French policymakers encouraged French bankers to make loans to the Russian government and opened negotiations for an entente. In 1891 a loose agreement provided for mutual consultation in crisis; in 1894 this was broadened into a military alliance by whose terms each partner promised to aid the other in case of attack by Germany or Germany's allies.

Franco-Russian Alliance

For a decade the Franco-Russian Alliance had little practical effect (though French loans did continue to flow to Russia). French diplomats turned to winning the Italians away from the Triple Alliance, and a Franco-Italian secret agreement in 1902 substantially weakened that Italian commitment. Of more central importance throughout the 1890s was recurrent tension between France and Britain, who had been at odds in various parts of the world and whose colonial competition at times seemed to threaten war. Britain's Boer War in South Africa added further ill feeling, and some British leaders began to urge an end to "splendid isolation" in favour of an entente with a continental power—most probably Germany, which was seen as part of an Anglo-Saxon racial bloc. But the German government responded coolly to overtures in this direction, thus feeding the fears of British leaders who saw Germany as a threat to British interests. The British turned to France instead and found a willing partner in the foreign minister Théophile Delcassé. A visit to Paris by King Edward VII in 1903 helped pave the way to the Anglo-French Entente Cordiale of 1904, which resolved all outstanding colonial conflicts between the two powers but stopped short of military alliance. The new entente was consolidated a year later, when French moves to take over Morocco as a protectorate were resented by the Germans, who thought they saw an opportunity to break up the new entente. Emperor William II offered Germany's support to the sultan of Morocco; this action irritated the British and led them to promise France strong support. In the conference of powers that followed at Algeciras (1906) France had to be content with special privileges rather than a protectorate in Morocco; but the Entente Cordiale was solidified, and it was Germany that thenceforth began to complain of isolation.

Stable government of the Bloc Républicain

The prewar years. From 1899 to 1905 a fairly coherent coalition of left-wing and centre parties (the so-called Bloc Républicain) provided France with stable government. The Cabinets headed by Waldeck-Rousseau in 1899–1902 and Émile Combes in 1902–1905 managed to liquidate the Dreyfus affair and to carry through the anticlerical reforms that culminated in the separation of church and state. The Entente Cordiale and the Russian alliance ensured France a more influential voice in European affairs. France possessed a colonial empire second only to Britain's in size. A new period of economic growth set in after the mid-1890s. Not surprisingly, later generations were to look back on the pre-1914 decade as *la belle époque* ("the good old days").

Still, some sources of sharp dissatisfaction and conflict remained. Many Roman Catholics were outraged by the triumph of the anticlericals, and they responded to the Vatican's urging to sabotage the new system. They resisted (sometimes violently) the transfer of church property to state ownership and refused to establish lay associations to govern the church. By 1907, however, resistance was clearly futile, and they began to accept the separation law as an accomplished fact. A difficult period followed for the church. The recruitment of priests fell off sharply, and many Catholic schools were closed for lack of funds. In the long run, however, the separation law reduced the intensity of conflict between Catholics and anticlericals. There was less reason for republicans to suspect and denounce a disestablished church.

A vocal minority on the right remained unreconciled to the radical republic and rallied round the banner of the Action Française, headed by Charles Maurras. This organization had developed at the height of the Dreyfus affair as a focal point for intellectuals who opposed a new trial for Dreyfus. Maurras, an aspiring young writer from the south, quickly emerged as its theorist and leader. In his view, France had gone astray in 1789 and had since been dominated by the "four alien nations"—Jews, Freemasons, Protestants, and *métèques* ("aliens"). He preached a return to stable institutions and an organic society, in which the monarchy and the church would be essential pillars. Maurras appealed to many traditionalists, professional men, churchmen, and army officers. Action Française readily resorted to both verbal and physical violence, and its organized bands, the Camelots du Roi, anticipated the tactics of later Fascist movements. By 1914 Maurras' movement, though still relatively small, was the most coherent and influential enemy of the republic.

Alienation of the working class

Equally serious was the alienation of much of the working class. The CGT remained officially committed to revolutionary syndicalism; it rejected political action as a useless diversion of the proletariat's energies and exalted the idea of the general strike as the proper weapon to destroy bourgeois society. Although the CGT attracted only about 10 percent of French workers (most workers stubbornly refused to join any union), it was aggressive enough to cause sporadic turmoil during the period 1906–10. Several major strikes were broken by forcible repression; the government either called out troops or mobilized the strikers (who were also reservists) into the army. Proposals for labour-reform legislation drew little support in a parliament dominated by representatives of the bourgeoisie and the peasantry.

Despite the CGT, most workers by now were voting for the new unified Socialist Party. But the SFIO refused to permit its deputies to participate in or support bourgeois Cabinets (a policy dictated to the French party in 1904 by the Second International, dominated by the German socialists) and thus condemned itself to an oppositionist stance in parliament. This destroyed the left-wing coalition that had given France stable Cabinets from 1899 to 1905. Socialist strength continued to rise, and by 1914, the party was second only to the Radicals in the Chamber of Deputies. Although its doctrine remained rigorously Marxist, in deference to the instructions of the International, the party's conduct was much more flexible. Jaurès, whose "humanitarian" socialism was in large part derived from an older French heritage of left-wing thought, guided the socialists in parliament toward informal cooperation with the bourgeois left in an effort to achieve domestic social reforms and an internationalist, antimilitarist foreign policy. Jaurès' central concern during the pre-1914 decade was to avert the general war that he saw looming ahead in Europe.

The socialist withdrawal from the Bloc Républicain in 1905 forced the Radicals to look to the other centre parties

Centre coalition governments

as coalition partners. Until 1914—and, indeed, most of the time until 1940—France was governed by heterogeneous centre coalitions in which the Radicals most often held the key posts. In 1906 the Radical Georges Clemenceau began a three-year premiership. He proposed a long list of social reforms, including the eight-hour day and an income tax, but parliament blocked virtually all of them. More surprising was Clemenceau's ruthless suppression of strikes and his vigorous, nationalistic foreign policy. In 1907 his government sponsored a rapprochement between Britain and Russia that completed the triangle of understandings thenceforth called the Triple Entente. But Clemenceau refused to risk war through all-out support of his Russian ally during the Bosnian crisis of 1908. When his Cabinet fell in 1909, Clemenceau had effectively alienated his own Radical Party and seemed unlikely ever to return to high office.

Clemenceau's successors, Aristide Briand and Joseph Caillaux, undertook a policy of détente in European affairs. Briand, like Clemenceau, belied his left-wing origins by forcibly repressing a major strike in 1910; in foreign affairs, however, he preferred a policy of coexistence with Germany. Caillaux pushed this latter experiment even further. In 1911 he had to deal with a new crisis in Morocco, where the French were again pushing toward a protectorate against German objections. When the Germans sent a gunboat to Morocco, Caillaux made an effort at appeasement, handing over to Germany a slice of the Congo as compensation. French patriots were outraged; the Caillaux Cabinet was overthrown and replaced in January 1912 by one headed by Raymond Poincaré.

Nationalist revival

There were signs of a changing intellectual mood in the country, especially among young Frenchmen. A nationalist revival affected many Frenchmen who for a decade had grown increasingly anxious about what they regarded as the puzzling and threatening attitude of Germany's post-Bismarckian leadership; they looked once more to the army as the nation's bulwark, and its prestige was on the rise. These nationalist tendencies found their embodiment in Poincaré, whose intransigent patriotism and determination to stand up to Germany were beyond doubt. As premier in 1912–13 Poincaré devoted himself to strengthening the armed forces and to reinvigorating France's alliance system. An agreement with the British provided for a new sharing of naval responsibilities: the French concentrating in the Mediterranean, the British in the North Sea. Poincaré made a state visit to Russia to revive the sagging Franco-Russian Alliance. In January 1913 he was elected to the presidency of the republic, where, he believed, he could ensure continuity of policy during his seven-year term. In 1913 the size of the standing army was increased by lengthening the conscription period from two to three years.

Poincaré found bitter opposition on the left. The socialists were strongly antimilitarist and hoped for an eventual reconciliation with Germany via collaboration between the two socialist parties. They clung to the belief that the working class everywhere could block war by resorting to a general strike. A large segment of the Radical Party followed the Caillaux line, favouring Franco-German collaboration through such ventures as banking consortia for joint investment abroad. Much of rural France also lacked enthusiasm for the new nationalistic mood. The combined strength of this opposition was revealed in the parliamentary elections of 1914, when the parties of the left won a narrow victory.

World War I. Before a change in policy could be imposed, however, a new crisis in the Balkans threatened a general war. The assassination of the Austrian archduke Francis Ferdinand in Sarajevo on June 28, 1914, inaugurated five weeks of feverish negotiations, in which France's role has been much debated. Some historians have accused Poincaré and his supporters of a willingness to go to the brink of war rather than seek a negotiated settlement or use restraint on the Serbs and Russians; Poincaré's state visit to St. Petersburg at the height of the crisis has been seen as an occasion for a French promise of full support to Russia. A more judicious view is that many French statesmen had long seen the possibility and even the likelihood of a general war, and they suspected that the German government desired such a war; the Poincaré group believed that under these circumstances France could not risk the loss of its allies. French support of the Serbs and the Russians, according to this view, was thus inspired by a calculated judgment regarding French security.

German declaration of war against France

Germany's declaration of war against France on August 3 produced a spontaneous outburst of patriotic sentiment. Trade union and Socialist leaders, some of whom had been on a governmental list of dangerous subversives to be arrested in case of war, rallied to the colours. A national union Cabinet was formed. Parliament, after voting war credits, went into an extended recess, handing over the conduct of the war to the Cabinet and the high command. During the initial months the high command made most of the crucial decisions; the Cabinet accorded almost unlimited freedom of action to the commander in chief, General Joseph Joffre, assuming that the war would last only a few weeks and that civilian interference would only prolong hostilities.

Joffre's war plans for an immediate advance across the frontier into the lost provinces of Alsace and Lorraine were suspended when German forces struck through Belgium and threatened late in August to envelop Paris. Joffre managed to blunt the German attack and force the Germans to more defensible positions. The rival armies dug into trench positions that remained largely static until 1918. Meanwhile, the French high command continued to believe that the fate of France would be decided on the Western Front. In 1916 a powerful German artillery attack on the French fortress positions surrounding Verdun lasted from February to June and cost each side some 250,000 men. For the French, the hero of Verdun was the sector commander, General Philippe Pétain.

Joffre was by now under heavy criticism in Paris. Both the Cabinet and the Chamber were determined to assert greater control over the war effort, so that the high command's authority was steadily whittled away. Joffre was finally replaced in late 1916 by General Robert Nivelle. All through 1917, rival factions in the Chamber debated the conduct of the war, backing different generals and threatening Cabinet crises. Worse still, morale among the troops reached a dangerous low point in 1917, culminating in serious mutinies that affected 54 French divisions. Pétain, who replaced Nivelle in May, managed to achieve stability by a judicious combination of severity and concessions.

Nevertheless, by autumn 1917 there was widespread defeatism in France and much talk of a "white peace." The Radical leader Caillaux was prepared to try for negotiations with the Germans; but his chance never came. When the Cabinet of Premier Paul Painlevé was overthrown in November 1917, President Poincaré recalled Georges Clemenceau to the premiership. Clemenceau stood for a fight to the finish. At age 76 he still had enormous energy and doggedness, and he infused a new spirit into the country. In March 1918, when the Germans launched a last major offensive in the West, Clemenceau replaced the cautious and pessimistic Pétain with a more attack-minded general, Ferdinand Foch, and persuaded the British as well to accept him as supreme commander. The German drive was checked. On November 11 an armistice was signed in Foch's railway car near Compiègne.

Effects of the war

The victory was won at enormous cost for France. Of the 8 million Frenchmen mobilized, 1.3 million had been killed and almost 1 million crippled. Large parts of northeastern France, the nation's most advanced industrial and agricultural area, were devastated. Industrial production had fallen to 60 percent of the prewar level; economic growth had been set back by a decade. The enormous cost of the war seriously undermined the franc and foreshadowed many years of currency fluctuation. Even deeper, though largely hidden, were the psychological lesions caused by the strain of protracted warfare and by the sentiment that France could not again endure such a test.

At the Paris Peace Conference in 1919, Clemenceau, as the principal French negotiator, declared that his goal was to ensure the nation's security against a renewed German aggression. He sought, therefore, to reduce Germany's power in every possible fashion and to hedge

Germany about by strong barrier nations. He knew, however, that France could not dictate the peace terms and that he would have to compromise with the Americans and British, to whom he looked for aid in case of German resurgence. His stubborn defense of French demands irritated France's wartime allies; but his willingness to compromise in the end alienated many Frenchmen, who charged him with sacrificing the nation's security. The critics—who included Poincaré and Foch—were particularly outraged when Clemenceau abandoned his initial demand that Germany give up all territory west of the Rhine and that the Saar basin be annexed to France. These and other concessions led many right-wing deputies to oppose the Treaty of Versailles when it was presented for ratification in the autumn of 1919. Joining the opposition were the Socialists, who argued that the treaty was too harsh and that democratic Germany should not be punished for the sins of the kaiser. A majority of the Chamber, however, reluctantly ratified Versailles and vowed to assure its enforcement to the letter.

The interwar years. Frenchmen concentrated much of their energy during the early 1920s on recovering from the war. The government undertook a vast program of reconstructing the devastated areas and had largely completed that task by 1925. To compensate for manpower losses, immigration barriers were lowered, and two million foreign workers flooded into the country. Underlying all other concerns, however, was anxiety about the nation's security and about financing the costs of war and reconstruction. The peace settlement, in the eyes of many Frenchmen, had not provided adequate guarantees; and except among Socialists and Radicals there was little confidence in the League of Nations. American and British promises to aid France in case of future attack had been written into the treaty, but they became meaningless when the United States Senate rejected Versailles.

German reparations. A clause in the treaty had ascribed war guilt to the Germans and their allies and had obligated Germany to make reparations; the total sum due was calculated in 1921 at $33 billion, but the French were aware that the British hoped to see this total reduced.

Bloc National

The general elections of November 1919 resulted in a massive majority for the right-wing coalition called the Bloc National. The new Chamber set out to enforce Versailles to the letter; it also sought traditional security guarantees, maintaining the largest standing army in Europe and attempting to encircle Germany with a ring of military allies (Belgium and Poland in 1920–21; Czechoslovakia, Romania, and Yugoslavia in 1924–27). But the central issue was that of German reparations. By the end of 1921 the British clearly favoured a reduction of the burden in order to get Germany back on its feet; this issue caused increasing strain between the British and French governments. Premier Briand, who seemed willing to compromise, was overthrown by the Chamber and replaced by the more intransigent Poincaré. Repeated German defaults on reparations deliveries led Poincaré in January 1923 to send French troops and engineers (supported by a token force of Belgians) into the Ruhr valley to force German compliance or, if necessary, to collect reparations by direct seizure. The German government attempted passive resistance but finally had to comply. Germany agreed in 1924 to a revised reparations settlement, the so-called Dawes Plan, and the French occupation forces were withdrawn. The plan enabled the Germans to meet their obligations on schedule during the rest of the decade, though only with the help of large American loans. In 1926 France and the United States finally reached agreement on another nagging problem—the repayment of French war debts for wartime deliveries of American munitions and other supplies.

Financial crisis. The aftermath of the Ruhr occupation was to cast doubt on its apparent success. The German republic was weakened by the runaway inflation of 1923 and its future clouded; the occupation had embittered Britain and the United States. Even among Frenchmen, the victory had left a sour aftertaste because the costs of the occupation forced an increase in French taxes. In the elections of 1924 Poincaré's Bloc National was beaten by a coalition of the left, the Cartel des Gauches, and the Radicals were returned to power. But their triumph was brief; they were confronted by the nation's worst financial crisis since the war. The shaky franc went into rapid decline until there seemed to be danger of complete financial collapse. Seven Cartel Cabinets in 1924–26 wrestled ineffectively with the problem; at last the Cartel gave up, and Poincaré returned. The latter's reputation for decisive character and conservative views enabled him to win the bankers' support and to embark on such measures as slashing government expenses and increasing taxes. The franc began to rise, and it finally stabilized at about one-fifth of its 1914 value. Poincaré was hailed as "saviour of the franc," and, when he resigned in 1929 for reasons of health, he was acclaimed as one of the Third Republic's outstanding statesmen.

Collective security. Poincaré, in his final term of office (1926–29), retained as foreign minister Aristide Briand, who had been named to that post by the Cartel in 1925 and who was to remain there for seven years almost without interruption. Briand sensed a change in the public mood after the Ruhr episode and proclaimed himself "the apostle of peace"; he formulated a policy that he called *apaisement.* His goal was to work for collective security through the League of Nations, for disarmament, and for a reconciliation with those Germans who favoured peaceful and cooperative methods. Briand found a ready partner in Gustav Stresemann, the German foreign minister. By the Pact of Locarno (1925), the French and German governments bound themselves not to use force to alter the existing Franco-German frontier. In subsequent years, France sponsored Germany's entry into the League of Nations and made a series of concessions softening various aspects of the Treaty of Versailles. A revised reparations agreement in 1929 (the Young Plan) further eased Germany's obligations, and in 1930 the French ended their occupation of the German Rhineland five years ahead of schedule.

Internal conflict on the left. Throughout the 1920s, however, much of the working class remained alienated from a regime that showed little concern for social reform. The CGT had emerged from the war with redoubled strength and energy, its membership swelled by the workers who had poured into new war industries in the Paris region. The Clemenceau government had rewarded labour for its war effort by legislating the eight-hour day in 1919; but when the unions pushed for more reforms, a deadlock ensued. An attempted general strike in May 1920 was easily broken, and thousands of discouraged and embittered workers abandoned the CGT. Labour's strength was further dissipated by the formation of rival Catholic and Communist trade-union federations in 1919 and 1921.

Socialist split

The political influence of the workers was further impaired by a split in the Socialist Party in 1920. During the war, Socialist opposition to the slaughter had become increasingly vocal. The Bolshevik Revolution in Russia had reinforced this trend and offered a model that attracted many French Socialists. From 1918 onward, conflict intensified among Socialists over the possibility of joining Lenin's Comintern (Third International). At the party's annual congress in Tours in December 1920 Lenin's partisans carried the day by a large majority and shortly renamed their organization the French Communist Party (SFIC). The minority, headed by Jaurès's disciple Léon Blum, walked out of the congress and retained the traditional name SFIO. Throughout the 1920s antagonisms between these two Marxist factions hampered the left and prevented workable coalitions. Neither the Socialists nor the Communists would enter bourgeois-dominated Cabinets; the Communists refused even to make electoral agreements in support of a single left-wing candidate. The trend through the 1920s was favourable to the Socialists, while the Communists steadily lost influence and members; in 1928 the Socialists won 107 seats in the Chamber, the Communists only 11. Many French Communists resented dictation from Moscow, and the decade saw a long series of resignations and purges; by 1930 the remnant had been thoroughly "bolshevized" on the pattern of Lenin's own party.

The Great Depression and political crises. France at the end of the 1920s had apparently recovered its prewar stability, prosperity, and self-confidence. For a time it even seemed immune to the economic crisis that spread through Europe beginning in 1929; France went serenely on behind its high-tariff barrier, a healthy island in a chaotic world. By 1931, however, France in its turn succumbed to the effects of the depression, and the impact was no less severe than elsewhere.

In 1932 the right-wing parties lost control of the Chamber to the Radicals and Socialists. The Radical leader Édouard Herriot returned to the premiership, with Socialist support but not participation. During the next two years Herriot and a series of successors groped for a solution to the deepening crisis. French nervousness was increased by the surge of Nazi power across the Rhine, culminating in Adolf Hitler's accession to the chancellorship in January 1933. Right-wing movements in France—some openly fascist, others advocating a more traditional authoritarianism—grew in size and activity. By 1934 the shaky coalition was at the mercy of an incident—the Stavisky scandal, a sordid affair that tarnished the reputations of several leading Radicals. Antiparliamentary groups of the far right seized the occasion to demonstrate against the regime; on February 6 a huge rally near the Chamber of Deputies degenerated into a bloody battle with armed police, during which 15 rioters were killed and 1,500 injured. Premier Édouard Daladier, confronted by a threat of civil war, resigned in favour of a national union Cabinet under former president Gaston Doumergue. The regime survived the crisis, but serious stress persisted. Right-wing agitation was countered by unity of action on the left, grouping all the left-wing parties and the CGT; even the Communists participated in this effort, which culminated in 1935 in the formation of the Popular Front.

Right-wing agitation

Doumergue's government had meanwhile disintegrated when Radical ministers resigned over the premier's increasingly authoritarian tone. Doumergue was soon replaced by Pierre Laval, a former socialist who had migrated toward the right. Laval embarked on a vigorous but unpopular attempt to combat the depression through traditional techniques: sharp cuts in government spending and increased taxes. These policies wrecked his Cabinet early in 1936 and became campaign issues in the parliamentary election that spring. That election, probably the most bitterly contested since 1877, gave the Popular Front a narrow majority of the popular vote and a large majority in the Chamber. The Socialists for the first time became the largest party; but the greatest proportional gain went to the Communists, who jumped from 10 to 72 seats.

Léon Blum, the Socialist leader, became premier. Blum, an intellectual, was the first French premier of Jewish origin. His ministers were mostly Socialists and Radicals; the Communists refused his urgent invitation to participate. At the very outset, a wave of sit-down strikes spread throughout the country, expressing workers' pent-up resentment toward past governments and their determination to get what they considered to be justice. Blum persuaded industrial leaders to grant immediate wage increases, which ended the strike. Then he pushed additional reforms through parliament: the 40-hour week, paid vacations, collective bargaining, and the seminationalization of the Bank of France. Many other reform bills, however, were stalled in committee or in the Senate, which remained much more conservative than the Chamber.

Blum's social reforms were costly and controversial and were not buttressed by a program of economic reforms that might have stimulated production and restored confidence. Production surged briefly, then lagged again; unemployment remained high, rising prices offset wage gains, a flight of capital set in. When Blum attempted to impose exchange controls, the Senate rebelled and overthrew his Cabinet (June 1937). The Popular Front held together for another year, but the Socialists and Radicals were irretrievably divided on economic policy. In April 1938 France returned once more to the usual pattern of homogeneous centre coalitions, with the Socialists in opposition. The Radical Daladier served as premier in 1938–40; his finance minister, Paul Reynaud, suspended most of the Popular Front reforms and sought economic recovery through more orthodox policies favoured by business.

German aggressions. Meanwhile, Hitler's accession had placed French governments in an increasingly grave foreign-policy dilemma. By 1934 many French leaders believed that a return of "Poincarism" was in order, and Doumergue's foreign minister, Louis Barthou, set out to reinforce and extend France's alliance system. He reaffirmed French ties with Poland and the "Little Entente" countries and sought new understandings with both Italy and the Soviet Union. Barthou's assassination in late 1934 weakened the new alliance policy, though Pierre Laval in 1935 paid visits to both Rome and Moscow and actually signed a mutual assistance treaty with the U.S.S.R.

Reinforcement and extension of alliances

Mussolini's invasion of Ethiopia in late 1935 and Hitler's military reoccupation of the Rhineland in March 1936 were serious blows to French policy. After consulting the British, the French Cabinet decided not to risk a confrontation with Hitler, who thus won a major diplomatic victory. Hitler promptly fortified the Rhine frontier, so that French guarantees of military aid to eastern European allies lost much credibility. Furthermore, Hitler and Mussolini joined forces against the status quo powers. With Italy lost, Frenchmen of the centre and right grew cool toward closer ties with the Soviet Union; they had counted on Italy to counterbalance Soviet influence. France found itself dangerously isolated, dependent on the small eastern European countries and on the uncertain prospect of British military support in crisis. Not surprisingly, French policy after 1936 showed signs of weakness and drift.

The outbreak of the Spanish Civil War in July 1936 posed a severe problem of conscience for Blum's Popular Front government: whether to send aid to the Spanish republic, the only other Popular Front regime in Europe. Reluctantly, Blum remained aloof; his Radical allies strongly opposed intervention and threatened to bring down the Cabinet. A new crisis developed in March 1938, when Hitler's troops for the first time crossed a frontier—into Austria. The French and British confined themselves to formal protests. German pressure on Czechoslovakia followed. Although France was formally committed to aid Czechoslovakia in case of aggression, Premier Daladier succumbed to British pressure to appease Hitler by a compromise settlement. The Munich Agreement of September 30 provided a breathing space but caused sharp dissension and self-doubt in France. When Hitler occupied what was left of Czechoslovakia in March 1939, it appeared to be too late for successful diplomatic or military resistance to Hitler, yet a failure to resist would hand over the Continent to German domination. From April until August the French and British sought to bring the Soviet Union into a joint pact against Hitler, with the French pressing the reluctant British to take the risks involved. A Soviet decision to break off negotiations and to sign a pact with Hitler instead was the last in a long chain of disasters for France. On September 3, two days after Germany invaded Poland, the French and British governments reluctantly declared war on Germany.

War against Germany

French and British attempts to aid the Poles would have been ineffective even if tried. Hitler's offer of peace immediately after Poland fell was rejected by the Western Allies. The Nazi armies smashed through The Netherlands and Belgium on May 10, 1940, and soon broke the French defensive lines near Sedan. The German blitz brought chaos all along the Allied front. In Paris, Premier Paul Reynaud (who had replaced Daladier in March) pleaded for emergency aid from Britain and the United States; the British sent some additional air units but were unwilling to denude their island of all air defense; U.S. President Franklin D. Roosevelt offered moral encouragement but not open intervention.

On June 10, with the Germans approaching Paris, the government departed for Tours and declared Paris an open city. British Prime Minister Winston Churchill twice flew to Tours in an effort to keep France in the war. But Reynaud, who favoured continued resistance (from North Africa, if necessary) rapidly lost ground to the defeatists in his Cabinet, headed by Pétain. On June 14 the Cabinet left Tours for Bordeaux. Churchill, in a last desperate

Surrender to Germany

effort, proposed a pact of "indissoluble union" that would merge France and Britain as a single nation. By the time the proposal reached Bordeaux on June 16, however, the Pétain faction had gotten control of the Cabinet. Reynaud resigned that evening; Pétain was appointed in his place and asked Germany for surrender terms. On June 22 an armistice was signed with the Germans, near Compiègne, in the same railway car that had been the scene of Foch's triumph in 1918. The armistice provided for the maintenance of a quasi-sovereign French state and for the division of the country into an occupied zone (northern France plus the western coast) and an unoccupied southern zone. France was made responsible for the German army's occupation costs. The French army was reduced to 100,000 men and the navy disarmed in its home ports.

Society and culture under the Third Republic. Under the Third Republic the middle and lower sectors of society came to share political and social dominance with the rich notables. Universal suffrage gave them a new political weapon; France's peculiar socioeconomic structure gave them political weight.

Economy. Republican France remained a nation of small producers, traders, and consumers. The surge of industrialization that marked the era of Napoleon III had stopped short of a full-scale industrial revolution. The new dynamic sector of the economy was far outweighed by a static or slowly changing sector. The bulk of industry remained smaller and more dispersed than in other industrializing countries. As late as the decade before 1914, 90 percent of France's industrial enterprises employed fewer than five workers each; in the extensive textile and clothing trades more than half of the employees still worked at home rather than in factories. Commerce and trade followed the same pattern, with small shops and banks surviving in profusion. Similarly, rural France was dominated by small, subsistence family farms. The proportion of farmers in the total active population, which stood at 52 percent in 1870, was still about 45 percent in 1914 and 35 percent in 1930. When grouped together, the small independent producers, traders, and farmers far outnumbered any other segment of society, including the proletariat.

The reasons given for this slow pace of socioeconomic change are varied: shortages of basic natural resources, a tradition of specialization in luxury items, a code of mores that emphasized prudent management rather than risky experiment and that regarded as ideal the "family firm," small enough to be financed and managed by the owners alone. In any case, French industrialization took a different form from that of England or Germany. An initial burst of growth in the 1850s was followed by several decades of much more gradual expansion, which did not threaten the existing structure of society and the underlying value system. Most segments of society were reasonably satisfied and felt no threat to their way of life (only the members of the working class, both urban and agricultural, considered themselves outsiders and victims rather than participants); thus, the stability of the system was ensured. Not until well into the 20th century, and especially after 1918, did this state of affairs begin to change.

Laissez-faire policy

The governments of the Third Republic were representative of the small independents and responsive to their interests. Most of the bourgeoisie and the peasantry wanted a laissez-faire policy: low taxes, hands off the affairs of private citizens. There was little popular enthusiasm for costly ventures in foreign policy or expensive social reforms; the major exception—the conquest of colonial empire—had to be accomplished somewhat secretively and with limited resources. Only in tariff policy was laissez-faire flagrantly violated by the government, with the active consent of its bourgeois supporters. When the low-tariff treaties of Napoleon III expired in 1877, the government promptly returned to protectionism. Much of French agriculture and industry was thereby protected against more efficient foreign producers and insulated against the need for modernization. The short-range interests of the small, independent producer were thus guaranteed; the prospect of harm to his longer-range interests—as well as to those of the nation as a whole—was not yet clear.

From 1873 to the mid-1890s the French economy experienced a period of slackness. This trend reflected a condition affecting most of Europe, although France suffered a special blow when an epidemic of phylloxera in 1875–87 destroyed one-third of the nation's vineyards. From 1896 to 1914 industrial output rose impressively, exports increased by 75 percent, prices returned to the pre-slump level; this upturn was also generally Europewide rather than peculiar to France; but some special factors, such as the opening of a vast new iron-ore field in French Lorraine, did increase the French rate of industrial expansion. By 1914 French Lorraine had become the major centre of French iron and steel production, and France had become the world's largest exporter of raw iron ore (primarily to Germany). Yet the French were being outpaced by rivals. In 1870 France had still ranked as the world's second industrial and trading nation; by 1914 it had fallen to a poor fourth. Much of the liquid capital that might have been used for business expansion at home was being siphoned off into foreign investment; by 1914 almost one-third of such available French capital had been placed abroad—one-fourth of that sum in Russia and only one-tenth in the French colonies. Yet few Frenchmen had serious doubts about the course of economic policy under the Third Republic.

Economic effects of World War I

Only after the war, and particularly after 1930, were such doubts widely shared. The disruptive impact of the war exceeded the understanding not only of most citizens but also of most political leaders. Efforts to return to normality were futile because the postwar world and France had changed vastly. The enormous cost of a four-year mobilization, of reconstruction, and of war debts had to be borne. By the time of the Great Depression the government had been forced to write off a large share of war costs by devaluing the franc (1928) to one-fifth of its old value, costing many Frenchmen on fixed incomes much of their savings and shaking their confidence in the future. Still, no large group of embittered déclassés was created, ripe for the appeals of a demagogue. And after 1926 there was a brief resurgence of prosperity, so that by the end of the decade the indexes of industrial production, foreign trade, and living standards had risen well above the 1914 peak. Some illusions about the future and hopes of a happy return to prewar stability could therefore be retained.

But by 1935 industrial production had fallen to 79 percent of the 1928 level and exports to 55 percent. Registered unemployment hovered at less than 500,000, but this figure concealed the fact that many urban workers were subsisting on family farms owned by relatives. Besides, the French exported much of their unemployment; thousands of immigrant workers lost their work permits and had to return home. Not until 1938–39 did a measure of recovery set in, thanks to Reynaud's business-oriented policies, plus the stimulus of rearmament. By the time war broke out again, France had barely returned to the predepression level.

The workers, always outside the bourgeois consensus, were by now largely hostile to the system; most of the gains they had finally achieved in 1936 had quickly been snatched away again. But in addition many bourgeois Frenchmen now questioned the virtues of the traditional system. The 1930s therefore brought an intense fermentation of political and social thought; dozens of study groups and movements sprang up in Paris, seeking or preaching doctrines of drastic renovation and structures of government that might carry them out.

Positivism

Cultural and scientific attainments. The cultural climate of the later 19th century in France, as in the Atlantic world generally, was strongly marked by the current called positivism. The post-1848 generation looked with contempt on what it considered the excesses and the bad taste of the preceding Romantic era. A new interest in science and a new vogue of realism in literature and the arts prevailed during the Second Empire; it was best embodied in the novels of Gustave Flaubert and the paintings of Gustave Courbet. By the 1870s this mood had formed into what its advocates regarded as a coherent philosophical system, the content and label of which they borrowed from the French thinker Auguste Comte. These self-styled positivists placed their faith in science and reason as the path to inevitable progress, with only the remnants of superstition (surviving mainly in the church) still blocking the hopeful future. The positivist temper is manifest in the novels of Émile Zola and the paintings of Impressionists such as Édouard Manet, Claude Monet, Edgar Degas, and Pierre-Auguste Renoir.

The French also showed great creativity in pure science and made major discoveries in a wide variety of fields. Among the most notable figures were Louis Pasteur in medicine, Pierre and Marie Curie in physics, Marcelin Berthelot in chemistry, Henri Poincaré in mathematics, and Jean-Martin Charcot in psychopathology. In the social sciences the work of Gustave Le Bon and Émile Durkheim had a broad and enduring impact.

Although the positivist mood prevailed at least until World War I, it was contested by a rival current of thought that from the 1890s onward began to assert itself. To some sensitive people of artistic temperament the positivist outlook seemed arid and narrow, neglecting the emotional side of man. This was the view of the school of poets, including Paul Verlaine and Stéphane Mallarmé, who called themselves Symbolists. A remarkable group of composers carried the new, romantic mood into music: mainstream works by composers such as Jules Massenet, Georges Bizet, and Camille Saint-Saëns were followed by the more experimental compositions of Claude Debussy and Maurice Ravel.

Of significance was the growing influence after 1890 of such writers and thinkers as Paul Bourget, Maurice Barrès, and Henri Bergson. Bourget's novels challenged what he called "brutal positivism" and asserted such traditional values as authority, the family, and the established order. Barrès preached what he called "integral nationalism"; he called for a return to "the sources of national energy," which he found in historic institutions, the soil of the fatherland, and the solidarity between the living and the dead. The philosopher Bergson attacked scientific dogmatism and exalted humankind's nonrational drives—notably a creative force that he called élan vital, which he held distinguishes heroic individuals and nations from the plodding herd.

This new spirit had its parallel in political thought and action as well: in the syndicalist doctrines of Georges Sorel, in the activism of a minority in the labour movement, and in the resurgent nationalism that strongly affected many French young people in the years just before 1914. It also brought a return to the church and to an emotional patriotism. In the fine arts a new generation of painters abandoned both realism and Impressionism. These so-called Postimpressionists were moved by an intense subjectivism, an urge to express in various ways the artist's inner vision and deeper emotions. The changed mood was best embodied in the work of Paul Cézanne, Paul Gauguin, and the Dutch immigrant Vincent van Gogh.

Dadaism and Surrealism

The terrible strain and disillusionment of World War I weighed heavily on French cultural life during the interwar era, leading to the development of the literary and artistic movement called Dadaism. Its program of calculated nonsense was inspired by a deep revulsion against the insanity of war and the positivist view that the world had sense and meaning. Dadaism soon gave way, though, to the more durable Surrealist movement, whose principal theorist and founder was the poet André Breton. The declared goal of Surrealist writers and artists was to free man's unconscious impulses from the distorting controls of rational reflection; creativity, they said, came from deep nonrational drives.

A number of France's most notable writers, however, remained within the older humanistic tradition, yet they likewise reflected the doubts and neuroses of an age of crisis. Marcel Proust, whose massive multivolume novel *À la recherche du temps perdu* (*Remembrance of Things Past*) began to appear in 1913, used the stream-of-consciousness technique to probe, in minutely introspective fashion, into the recesses of his own mind and memory. André Gide, in similarly sensitive and introspective fashion, wrestled with the psychological difficulties arising from the conflict between a bourgeois society's values and the individual's instinctive drives.

As the mood of crisis deepened in the 1930s, so did the intensity of the challenge to old values, bringing men of frankly fascist temper, such as Robert Brasillach, and brutally nihilistic literary experimenters, such as Louis-Ferdinand Céline. But other Frenchmen continued to create works in the older tradition, including the social commentary of Roger Martin du Gard, Georges Duhamel, Jules Romains, and François Mauriac; the Neoromantic novels of André Malraux, preaching a modern gospel of heroic activism; the first writings of Jean-Paul Sartre; and the essays of Emmanuel Mounier, who was to inspire the new Catholic left after World War II.

FRANCE SINCE 1940

Wartime France. The German victory left the French groping for a new policy and new leadership. Some 30 prominent politicians—among them Édouard Daladier and Pierre Mendès-France—left for North Africa to set up a government-in-exile there; but Pétain blocked that enterprise by ordering their arrest on arrival in Morocco. The undersecretary of war in the fallen Reynaud Cabinet, General Charles de Gaulle, had already flown to London and in a radio appeal on June 18, 1940, summoned French patriots to continue the fight; but few heard or heeded his call in the first weeks. It was to Pétain, rather, that most of the nation looked for salvation.

The Vichy government. Parliament met at Vichy on July 9–10 to consider France's future. The session was dominated by Pierre Laval, Pétain's vice premier, who was already emerging as the strongman of the government. Laval, convinced that Germany had won the war and would thenceforth control the Continent, saw it as his duty to adapt France to the new authoritarian age. By skillful manipulation, he persuaded parliament to vote itself and the Third Republic out of existence. The vote (569 to 80) authorized Pétain to draft a new constitution. The draft was never completed, but Pétain and his advisers did embark on a series of piecemeal reforms, which they labeled the National Revolution. Soon the elements of a corporative state began to emerge, and steps were taken to decentralize France by reviving the old provinces. In the early stages of Vichy, Pétain's inner circle—except for Laval and a few others—was made up of right-wing traditionalists and authoritarians. The real pro-fascists, such as Jacques Doriot and Marcel Déat, who wanted a system modeled frankly on those of Hitler and Mussolini, soon left Vichy and settled in Paris, where they accepted German subsidies and intrigued against Pétain.

The end of the Third Republic

In December 1940 Pétain dismissed Laval and placed him briefly under house arrest. Laval had offended Pétain and his followers by his arrogance and his obvious taste for intrigue. His critics charged him also with attempting to bring Vichy France back into the war in alliance with the Germans. Both Laval and Pétain had accepted Hitler's invitation to a meeting at Montoire on Oct. 24, 1940, and during the weeks that followed the French leaders had publicly advocated Franco-German "collaboration." Whether Laval hoped for a real Franco-German alliance remains somewhat controversial. If so, it was a futile effort because Hitler had no interest in accepting France as a trusted partner; "collaboration" remained a French and not a German slogan. Hitler tolerated the temporary existence of a quasi-independent Vichy state as a useful device to help police the country and to collect the enormously inflated occupation costs levied by the armistice.

Laval was succeeded by another prewar politician, Pierre-Étienne Flandin, and he, in turn, by Admiral François Darlan, who was intensely anti-British and an intriguer by nature and who followed a devious path that involved continuing efforts at active collaboration with the Germans. Hitler, meanwhile, concentrated on draining France of raw materials and foodstuffs that were useful for the conduct of the war.

In April 1942 Pétain restored Laval to power, partly under German pressure. Laval retained that post until the collapse of Vichy in 1944. His role was increasingly difficult because the terrible drain of the war in the Soviet Union caused the Germans to increase their exactions. The Germans were short of manpower for their factories, and

Laval, under heavy pressure, agreed to the conscription of able-bodied French workers in return for the release of some French prisoners of war. He also assumed the task of repressing the French underground movement, whose activities hampered the delivery of supplies and men to Germany. After the war Laval and his friends were to argue that he had played a "double game" of limited collaboration to protect France against a worse fate.

German control of all France

Most of Vichy's remaining autonomy and authority was destroyed in November 1942, in direct consequence of the Anglo-American landings in North Africa. Vichy troops in Morocco and Algeria briefly resisted the American invasion, then capitulated when Admiral Darlan, who happened to be visiting Algiers at the time, negotiated an armistice. On November 11 Hitler ordered his troops in the occupied zone to cross the demarcation line and to take over all of France. The Vichy government survived, but only on German sufferance—a shadowy regime with little power and declining prestige.

The Resistance. Vichy's decline was paralleled by the rise of the anti-German underground. Within weeks of the 1940 collapse, tiny groups of French citizens had begun to resist. Some collected military intelligence for transmission to London; some organized escape routes for British airmen who had been shot down; some circulated anti-German leaflets; some engaged in sabotage of railways and German installations. The Resistance movement received an important infusion of strength in June 1941, when Hitler's attack on the Soviet Union brought the French Communist Party into active participation in the anti-German struggle. It was further reinforced by the German decision to conscript French workers; many draftees took to the hills and joined guerrilla bands that took the name maquis ("underbrush"). A kind of national unity was finally achieved in May 1943, when de Gaulle's personal representative, Jean Moulin, succeeded in establishing a National Resistance Council (Conseil National de la Résistance) that joined all the major movements into one federation.

De Gaulle's original call for resistance had attracted only a handful of French citizens who happened to be in Britain at the time. But, as the British continued to fight, a trickle of volunteers from France began to find its way to his headquarters in London. De Gaulle promptly established an organization called Free France and in 1941 capped it with a body called the French National Committee (Comité National Français), for which he boldly claimed the status of a legal government-in-exile. During the next three years, first in London and then (after 1943) in Algiers, he insisted on his right to speak for France and on France's right to be heard as a great power in the councils of the Allies. His demands and his manner irked Churchill and Roosevelt and caused persistent tension. The American government unsuccessfully attempted in 1942 to sidetrack him in favour of General Henri Giraud, who immediately after the Allied landings in North Africa was brought out of France to command the French armies in liberated North Africa and to assume a political role as well. De Gaulle arrived in Algiers in May 1943 and joined Giraud as copresident of a new French Committee of National Liberation. By the end of the year he had outmaneuvered Giraud and emerged as the unchallenged spokesman for French resisters everywhere. Even the Communists in 1943 grudgingly accepted his leadership.

Liberation. When the Allied forces landed in Normandy on June 6, 1944, the armed underground units had grown large enough to play an important role in the battles that followed—harassing the German forces and sabotaging railways and bridges. As the Germans gradually fell back, local Resistance organizations took over town halls and prefectures from Vichy incumbents. De Gaulle's provisional government immediately sent its own delegates into the liberated areas to ensure an orderly transfer of power. On August 19 Resistance forces in Paris launched an insurrection against the German occupiers, and on August 25 Free French units under General Jacques Leclerc entered the city. De Gaulle himself arrived later that day, and on the next he headed a triumphal parade down the Champs-Élysées. Most high-ranking Vichy officials (including Pétain and Laval) had moved eastward with the Germans; at the castle of Sigmaringen in Germany they adopted the posture of a government-in-exile.

Retribution against Vichy officials

De Gaulle's provisional government, formally recognized in October 1944 by the U.S., British, and Soviet governments, enjoyed unchallenged authority in liberated France. But the country had been stripped of raw materials and food by the Germans; the transportation system was severely disrupted by air bombardment and sabotage; 2.5 million French prisoners of war, conscripted workers, and deportees were still in German camps; and the task of liquidating the Vichy heritage threatened to cause grave domestic stress. An informal and spontaneous purge of Vichy officials or supporters had already begun in the summer of 1944; summary executions by Resistance bands appear to have exceeded 10,000.

A more systematic retribution followed. Special courts set up to try citizens accused of collaboration heard 125,000 cases during the next two years. Some 50,000 offenders were punished by "national degradation" (loss of civic rights for a period of years), almost 40,000 received prison terms, and between 700 and 800 were executed.

The Fourth Republic. Shortly after his return to Paris, de Gaulle announced that the citizens of France would determine their future governmental system as soon as the absent prisoners and deportees could be repatriated. That process was largely completed by midsummer 1945, soon after Germany's defeat, whereupon de Gaulle scheduled a combined referendum and election for October. Women, for the first time in French history, were granted suffrage. By an overwhelming majority (96 percent of the votes cast), the nation rejected a return to the prewar regime. The mood of the liberation cra was markcd by a thirst for renovation and for change.

New men of the Resistance movement dominated the constituent assembly, and the centre of gravity was heavily to the left: three-fourths of the dcputics werc Communists, Socialists, or Christian Democrats who had adhered to the new party of the Catholic left—the Popular Republican Movemcnt (Mouvcment Rcpublicain Populaire).

Constitution of the Fourth Republic. It soon becamc clear that the apparent unity forged in the Resistance was supcrficial and that the new political elitc was sharply divided over the form of the new rcpublic. Some urged the need for greater stability through a strong executive; others, notably the Communists, favoured concentrating power in a one-housc legislature subject to grass-roots control by the voters. De Gaullc rcmaincd aloof from this controversy, though it was obvious that he favoured a strong presidency. In January 1946 de Gaullc suddenly resigned his post as provisional presidcnt, apparently cxpecting that a wave of public support would bring him back to power with a mandate to impose his constitutional ideas. Instead, the public was stunncd and confused, and it failed to react. The assembly promptly chose thc Socialist Fèlix Gouin to replace him, and thc embittered dc Gaullc retired to his country estatc.

The assembly's constitutional draft, submitted to a popular referendum in May 1946, was rcjected by the voters. A new assembly was quickly elected to prepare a rcvised draft, which in October was narrowly approved by the voters. De Gaulle actively intervened in the campaign for the second referendum, denouncing the proposed systcm as unworkable and urging the need for a stronger executive. His ideas anticipated the systcm that later was to be embodied in the constitution of thc Fifth Republic (1958).

Political and social changes. The structure of the Fourth Republic seemed remarkably like that of the Third; in actual operation it seemed even more familiar. The lower house of parliament (now renamed the National Assembly) was once more the locus of power; shaky coalition Cabinets again succeeded one another at brief intervals, and the lack of a clear-cut majority in the country or in parliament hampered vigorous or coherent action. Many politicians from the prewar period turned up once again in Cabinet posts.

Yet outside the realm of political structure and parliamentary gamesmanship there were real and fundamental changes. The long sequence of crises that had shaken the nation since 1930 had left a deep imprint on French attitudes. There was much less public complacency; both the routines and the values of the French people had been shaken up and subjected to challenge by a generation of upheaval. Many of the new men who had emerged from the Resistance movement into political life, business posts, or the state bureaucracy retained a strong urge toward renovation as well as to a reassertion of France's lost greatness.

This altered mood helps to explain the economic growth that marked the later years of the Fourth Republic. After a painful period of slow recovery from the war, aided by massive economic aid from the United States, a burst of industrial expansion in most branches of the economy began in the mid-1950s, unmatched in any decade of French history since the 1850s. The rate of growth for a time rivaled that of Germany and exceeded that of most other European countries. The only serious flaw in the boom was a nagging inflationary trend that weakened the franc and undermined the competitiveness of French exports. Short-lived coalition Cabinets were incapable of taking the painful measures needed to check this trend.

Industrial expansion in the 1950s

Colonial independence movements. A less fortunate aspect of the national urge to reassert France's stature in the world was the Fourth Republic's costly effort to hold the colonial empire. France's colonies had provided de Gaulle with his first important base of support as leader of Free France, and as the war continued they had furnished valuable resources and manpower. The colonial peoples, therefore, now felt justified in demanding a new relationship with France, and French leaders recognized the need to grant concessions. But most of these leaders, including de Gaulle, were not prepared to permit any infringement on French sovereignty, either immediately or in the foreseeable future. For a nation seeking to rebuild its self-respect, the prospect of a loss of empire seemed unacceptable; most of the French, moreover, were convinced that the native peoples overseas lacked the necessary training for self-government and that a relaxation of the French grip would merely open the way to domination by another imperial power. The constitution of 1946 therefore introduced only mild reforms: the empire was renamed the French Union, within which the colonial peoples would enjoy a narrowly limited local autonomy plus some representation in the French parliament.

This cautious reform came too late to win acceptance in many parts of the empire. The situation was most serious in Southeast Asia, where the Japanese had displaced the French during World War II. Japan's defeat in 1945 enabled the French to regain control of southern Indochina, but the northern half was promptly taken over by a Vietnamese nationalist movement headed by the communist Ho Chi Minh. French efforts to negotiate a compromise with Ho's regime broke down in December 1946, and a bloody eight-year war followed. In the end, the financial and psychological strain proved too great for France to bear, and, after the capture of the French stronghold of Dien Bien Phu in 1954 by the Vietnamese, the French sought a face-saving solution. A conference of interested powers at Geneva that year ended the war by establishing what was intended as a temporary division of Vietnam into independent northern and southern states. Two other segments of Indochina, the former protectorates of Laos and Cambodia, had earlier been converted by the French into independent monarchies to preserve some French influence there.

Algerian revolt

On the night of Oct. 31, 1954, barely six months after the fighting in Indochina ended, Algerian nationalists raised the standard of rebellion. By 1958 more than a half million French soldiers had been sent to Algeria—the largest overseas expeditionary force in French history. France's determination to hold Algeria stemmed from a number of factors: the presence of almost a million European settlers, the legal fiction that Algeria was an integral part of France, and the recent discovery of oil in the southern desert. Fears that the rebellion might spread to Tunisia and Morocco led the French to make drastic concessions there; in 1956 both of these protectorates became sovereign states.

The long and brutal struggle in Algeria gravely affected the political life of the Fourth Republic and ended by destroying it. A vocal minority in France openly favoured a negotiated settlement, though no political leader dared take so unpopular a position. Right-wing activists, outraged at what they saw as the spread of defeatism, turned to conspiracy; both in Paris and in Algiers, extremist groups began to plot the replacement of the Fourth Republic by a tougher regime, headed by army officers or perhaps by General de Gaulle.

These plans had not yet matured when a Cabinet crisis in April–May 1958 gave the conspirators a chance to strike. On May 13, when a new Cabinet was scheduled to present its program to the National Assembly, activist groups in Algiers went into the streets in an effort to influence parliament's vote. By nightfall they were in control of the city and set up an emergency government with local army support. De Gaulle on May 15 announced that he was prepared to take power if called to do so by his fellow citizens. Two weeks of negotiations followed, interspersed with threats of violent action by the Algiers rebels. Most of the Fourth Republic's political leaders reluctantly concluded that de Gaulle's return was the only alternative to an army coup that might lead to civil war. On June 1, therefore, the National Assembly voted de Gaulle full powers for six months, thus putting a de facto end to the Fourth Republic.

The Fifth Republic. During his years of self-imposed exile, de Gaulle had scorned and derided the Fourth Republic and its leaders. He had briefly sought to oppose the regime by organizing a Gaullist party, but he had soon abandoned this venture as futile. Back in power, he adopted a more conciliatory line; he invited a number of old politicians to join his Cabinet, but he made sure that his own ideas would shape the future by naming his disciple Michel Debré head of a commission to draft a new constitution. This draft, approved in a referendum in September by 79 percent of the valid votes cast, embodied de Gaulle's conceptions of how France should be governed. Executive power was considerably increased at the expense of the National Assembly. The president of the republic was given much broader authority; he would henceforth be chosen by an electorate of local notables rather than by parliament, and he would select the premier (renamed prime minister), who would continue to be responsible to the National Assembly but would be less subject to its whims. In the new National Assembly, elected in November, the largest block of seats was won by a newly organized Gaullist party, the Union for the New Republic (Union pour la Nouvelle République; UNR); the parties of the left suffered serious losses. In December de Gaulle was elected president for a seven-year term, and he appointed Debré as his first prime minister. The Fifth Republic came into operation on Jan. 8, 1959, when de Gaulle assumed his presidential functions and appointed a new government.

Expanded executive power

The new president's most immediate problems were the Algerian conflict and the inflation caused by the war. He attacked the latter, with considerable success, by a program of deflation and austerity. As for Algeria, he seemed at first to share the views of those whose slogan was "Algérie française"; but, as time went by, it became clear that he was seeking a compromise that would keep an autonomous Algeria loosely linked with France. The Algerian nationalist leaders, however, were not interested in compromise, while the diehard French colonists looked increasingly to the army for support against what they began to call de Gaulle's betrayal. Open sedition followed in 1961, when a group of high army officers headed by General Raoul Salan formed the Secret Army Organization (Organisation de l'Armée Secrète; OAS) and attempted to stage a coup in Algiers. When the insurrection failed, the OAS turned to terrorism; there were several attempts on de Gaulle's life. The president pushed ahead nevertheless with his search for a settlement with the Algerians that would combine independence with guarantees for the safety of French colonists and their property. Such a settlement was finally worked out, and in a referendum (April 1962) more than 90 percent of the war-weary French voters approved the agreement. An exodus of European

Free Algeria

settlers ensued; 750,000 refugees flooded into France. The burden of absorbing them was heavy, but the prosperous French economy was able to finance the process despite some psychological strains.

The Algerian crisis speeded the process of decolonization in the rest of the empire. Some concessions to local nationalist sentiment had already been made during the 1950s, and de Gaulle's new constitution had authorized increased self-rule. But the urge for independence was irresistible, and by 1961 virtually all the French territories in Africa had demanded and achieved it. De Gaulle's government reacted shrewdly by embarking on a program of military support and economic aid to the former colonies; most of France's foreign-aid money went to them. This encouraged the emergence of a French-speaking bloc of nations, which gave greater resonance to France's role in world affairs.

The Algerian settlement brought France a respite after 16 years of almost unbroken colonial wars. Prime Minister Debré resigned in 1962 and was replaced by one of de Gaulle's closest aides, Georges Pompidou. The party leaders now began to talk of amending the constitution to restore the powers of the National Assembly. Faced by this prospect, de Gaulle seized the initiative by proposing his own constitutional amendment; it provided for direct popular election of the president, thus further increasing his authority. When de Gaulle's critics denounced the project as unconstitutional, de Gaulle retaliated by dissolving the assembly and proceeding with his constitutional referendum. On October 28, 62 percent of those voting gave their approval, and in the subsequent elections (November) the Gaullist UNR won a clear majority in the assembly. Pompidou was reappointed prime minister.

The Gaullist era

When de Gaulle's presidential term ended in 1965, he announced his candidacy for reelection. For the first time since 1848 the voting was to be by direct popular suffrage. De Gaulle's challengers forced de Gaulle into a runoff, and his victory over the moderate leftist François Mitterrand in the second round by a 55–45 margin was closer than had been predicted but sufficed to assure him of seven more years in power. Although de Gaulle's leadership had not ended political division in France, his compatriots could not ignore the achievements of his first term. Not only had he disengaged France from Algeria without producing a civil war at home, but he could also point to continuing economic growth, a solid currency, and a stability of government that was greater than any living French citizen had known.

The mid-1960s were the golden years of the Gaullist era, with the president playing the role of elected monarch and respected world statesman. France had adjusted well to the loss of empire and to membership in the European Common Market, which brought the country more benefits than costs. De Gaulle could now embark on an assertive foreign policy, designed to restore what he called France's *grandeur;* he could indulge in such luxuries as blocking Britain's entry into the Common Market, ejecting NATO forces from France, lecturing the Americans on their involvement in Vietnam, and traveling to Canada to call for a "free Quebec." He continued the Fourth Republic's initiative in developing both nuclear power and nuclear weapons—the so-called *force de frappe.* His foreign policy enjoyed broad domestic support, and the French people also seemed content with the prosperity and order that accompanied his paternalistic rule.

Beneath the surface, however, basic discontents persisted, and they were startlingly revealed by the crisis that erupted in May 1968. Student disorders in the universities of the Paris region had been sporadic for some time; they exploded on May 3, when a rally of student radicals at the Sorbonne became violent and was broken up by the police. This minor incident quickly became a major confrontation: barricades went up in the Latin Quarter, street fighting broke out, and the Sorbonne was occupied by student rebels, who converted it into a huge commune. The unrest spread to other universities and then to the factories as well; a wave of wildcat strikes rolled across France, eventually involving several million workers and virtually paralyzing the nation. Prime Minister Pompidou ordered the police to evacuate the Latin Quarter and concentrated on negotiations with the labour union leaders. An agreement calling for improved wages and working conditions was hammered out, but it collapsed when the rank-and-file workers refused to end their strike.

By the end of May various radical factions no longer concealed their intent to carry out a true revolution that would bring down the Fifth Republic. De Gaulle seemed incapable of grappling with the crisis or of even understanding its nature. The Communist and trade union leaders, however, provided him with breathing space; they opposed further upheaval, evidently fearing the loss of their followers to their more extremist and anarchist rivals. In addition, many middle-class citizens who had initially enjoyed the excitement lost their enthusiasm as they saw established institutions disintegrating before their eyes.

De Gaulle, sensing the opportune moment, suddenly left Paris by helicopter on May 29. Rumours spread that he was about to resign. Instead, he returned the next day with a promise of armed support, if needed, from the commanders of the French occupation troops in Germany. In a dramatic four-minute radio address, he appealed to the partisans of law and order and presented himself as the only barrier to anarchy or Communist rule. Loyal Gaullists and nervous citizens rallied round him; the activist factions were isolated when the Communists refused to join them in a resort to force. The confrontation moved from the streets to the polls. De Gaulle dissolved the National Assembly, and on June 23 and 30 the Gaullists won a landslide victory. The Gaullist Union of Democrats for the Republic (Union des Démocrates pour la République; UDR, the former UNR), with its allies, emerged with three-fourths of the seats.

The repercussions of the May crisis were nevertheless considerable. The government, shocked by the depth and extent of discontent, made a series of concessions to the protesting groups. Workers were granted higher wages and improved working conditions; the assembly adopted a university reform bill intended to modernize higher education and to give teachers and students a voice in running their institutions. De Gaulle took the occasion to shake up his Cabinet; Pompidou was replaced by Maurice Couve de Murville. De Gaulle evidently sensed the emergence of Pompidou as a serious rival, for the prime minister had shown toughness and nerve during the crisis, while the president had temporarily lost his bearings. The economy also suffered from the upheaval; austerity measures were needed to stabilize things once more.

Although normalcy gradually returned, de Gaulle remained baffled and irritated by what the French called *les événements de mai* ("the events of May"). Perhaps it was to reaffirm his leadership that he proposed another test at the polls: a pair of constitutional amendments to be voted on by referendum. Their content was of secondary importance, yet de Gaulle threw his prestige into the balance, announcing that he would resign if the amendments failed to be approved. Every opposition faction seized upon the chance to challenge the president. On April 27, 1969, the amendments were defeated by a 53 to 47 percent margin, and that night de Gaulle silently abandoned his office. He returned to the obscurity of his country estate and turned once more to the writing of his memoirs. In 1970, just before his 80th birthday, he died of a massive stroke. His passing inspired an almost worldwide chorus of praise, even from those who up to then had been his most persistent critics.

France after de Gaulle. De Gaulle's departure from the scene provoked some early speculation about the survival of the Fifth Republic and of the Gaullist party (the UDR); both, after all, had been tailored to the general's measure. But both proved to be durable, although his successors gave the system a somewhat different tone. Georges Pompidou won the presidency in June 1969 over several left and centre rivals. He adopted a somewhat less assertive foreign policy stance and in domestic affairs showed a preference for classical laissez-faire, reflecting his connections with the business community.

The turn toward a more conservative, business-oriented line contributed to a revival of the political left, which had been decimated by the aftershocks of the events of May 1968. François Mitterrand, leader of a small left-centre party, took advantage of the change in political climate. In 1971 he engineered a merger of several minor factions with the almost moribund Socialist Party and won election as leader of the reinvigorated party. He then persuaded the Communists to join the Socialists in drafting what was called the Common Program, which was a plan to combine forces in future elections and in an eventual coalition government.

The Common Program

Unexpectedly in April 1974 President Pompidou died of cancer. Mitterrand declared his candidacy as representative of the united left, while the conservatives failed to agree on a candidate. The Gaullists nominated Prime Minister Jacques Chaban-Delmas, but a sizable minority of the UDR broke ranks and instead declared support for a non-Gaullist, Valéry Giscard d'Estaing, who was the leader of a business party, the Independent Republicans (Républicains Indépendants). Giscard won over Chaban-Delmas in the first round and narrowly defeated Mitterrand in the runoff.

Despite his conservative connections, the new president declared his goal to be the transformation of France into "an advanced liberal society." As prime minister he chose the young and forceful Jacques Chirac, leader of the Gaullist minority that had bolted the UDR in Giscard's favour. The new leadership pushed through a reform program designed to attract young voters: it reduced the voting age to 18, legalized abortion within certain limits, and instituted measures to protect the environment. But the course of reform was stalled by the oil crisis of 1973, brought on by events in the Middle East. Industrial production slowed, unemployment rose, and inflation threatened.

As discontent grew, Giscard's leadership was challenged by his ambitious prime minister, Chirac. Open rivalry between the two men led Giscard to dismiss Chirac in favour of Raymond Barre, a professional economist. Chirac retaliated by persuading the divided and disheartened Gaullists to transform the UDR into a new party, the Rally for the Republic (Rassemblement pour la République; RPR), with himself as its head. He also gained an additional power base by standing successfully for election to the revived post of mayor of Paris.

While these factional conflicts on the right seemed to assure victory for the united left in the 1978 parliamentary elections, the Socialist-Communist alliance fell apart. The Socialists had made dramatic gains at Communist expense since the Common Program had been adopted, and the Communists decided it was safer to scuttle the agreement. The collapse of leftist unity alienated a large number of left voters and enabled the conservatives to retain control of the National Assembly.

Victory of the left

When Giscard's presidential term ended in May 1981, opinion polls seemed to indicate that he would be elected to a second term. He overcame a vigorous challenge by Chirac in the first round of voting and seemed well placed to defeat the Socialist Mitterrand in the runoff. But Mitterrand surprised the pollsters by scoring a slim victory—the first major victory for the left in three decades. Profiting from the wave of euphoria that followed, Mitterrand dissolved the National Assembly and succeeded once again. The Socialists won a clear majority of seats (269 of the total 491) and seemed in a position to transform France into a social democratic state.

France under a Socialist presidency. *Mitterrand's first term.* Mitterrand moved at once to carry out what appeared to be the voters' mandate. He named as prime minister a longtime Socialist militant, Pierre Mauroy, whose Cabinet was almost solidly Socialist except for four Communists. Major reforms followed quickly. A broad sector of the economy was nationalized (including 11 large industrial conglomerates and most private banks); administrative decentralization shifted part of the state's authority to regional and local councils; social benefits were expanded and factory layoffs made subject to state controls; tax rates were increased at the upper levels; and a special wealth tax was imposed on large fortunes.

The Socialists hoped that other industrial countries would adopt similar measures and that this joint effort would stimulate a broad recovery from the post-1973 recession. Instead, most of the other Western nations took the opposite course, turning toward conservative retrenchment. Isolated in an unsympathetic world and hampered by angry opposition at home, the Socialist experiment sputtered: exports declined, the value of the franc fell, unemployment continued to rise, and capital fled to safe havens abroad. The government was soon forced to retreat. Mauroy was replaced by a young Socialist technocrat, Laurent Fabius, who announced a turn from ideology to efficiency, with modernization the new keynote.

Many leftist voters were disillusioned by this shift. On the far left the Communists withdrew their ministers from the Cabinet. On the far right a new focus of discontent emerged in Jean-Marie Le Pen's National Front (Front National), which scored successes with its xenophobic campaign to expel immigrant workers. To nobody's surprise the Socialists lost control of the National Assembly to the conservative coalition in the March 1986 elections.

Cohabitation

Mitterrand's presidential term still had two years to run. But the Fifth Republic now faced a long-debated test: Could the system function when parliament and president were at odds? Mitterrand chose the path of prudent retreat. He named as prime minister the conservatives' strongest leader, Jacques Chirac of the Gaullist RPR, and abandoned to him most governmental decisions (except on foreign and defense policy, which de Gaulle himself had reserved for the president). This uneasy relationship was promptly labeled "cohabitation"; it lasted two years and in the end worked in Mitterrand's rather than Chirac's favour.

Chirac acted at once to reverse many of the Socialists' reforms. He began the complex process of privatizing the nationalized enterprises, reduced income tax rates at the upper levels and abolished the wealth tax, and removed some of the regulatory controls on industry. These moves brought Chirac praise but also criticism. His popularity suffered in addition from a series of threats to public order—notably a long transport strike and a wave of terrorist attacks on the streets of Paris—that cast some doubt on the government's promise to ensure law and order. As Chirac's approval ratings fell, Mitterrand's recovered. Cohabitation enabled him to avoid making sensitive decisions, and voters gave him credit for faithfully respecting his constitutional limitations.

Mitterrand's second term. Restraint paid dividends when Mitterrand ran for a second term in April–May 1988 and scored a clear victory against Chirac. The resurgent president chose the Socialist Michel Rocard as prime minister and once again dissolved the National Assembly in the hope that the voters would give him a parliamentary majority. However, the Socialists and their allies fell short of a clear majority.

Mitterrand's choice of Rocard as prime minister caused some surprise, for the two men had headed rival factions within the Socialist Party, and they were temperamentally alien. Rocard was a brilliant financial expert and an advocate of government by consensus of left and centre, while Mitterrand was considered a master of political gamesmanship. The uneasy relationship lasted three years, and Rocard managed the economy well enough to maintain his high approval rating in the polls until the end.

Mitterrand's decision to replace Rocard in 1991 with France's first woman prime minister, Edith Cresson, provoked serious controversy. Cresson, a Mitterrand loyalist, had held a variety of Cabinet posts during the 1980s and was seen as an able but tough and abrasive politician. Brash public statements hurt her approval rating, and, after the Socialists suffered disastrous losses in regional elections (March 1992), Mitterrand replaced Cresson with a different sort of Socialist, Pierre Bérégovoy.

Pierre Bérégovoy

Bérégovoy was a rare example of a proletarian who had risen through trade union ranks to political eminence. He had a reputation as an expert on public finance and as an incorruptible politician. His promise to end the plague of financial scandals that had beset recent Socialist governments won applause but left him vulnerable when he, in turn, was accused of having accepted a loan under compromising circumstances. Although no illegality was involved, Bérégovoy's reputation for integrity suffered. In the 1993 parliamentary elections, the Socialists suffered a crushing defeat by the right-wing coalition (RPR and UDR). Bérégovoy resigned as prime minister and a few weeks later shocked the country by committing suicide.

Although the triumphant conservatives called on Mitterrand to resign, he refused; his presidential term still had two years to run. But he had to face cohabitation again, this time with another Gaullist, Édouard Balladur. Chirac chose to avoid the risks of active decision making while he was preparing his own campaign for the presidency.

Mitterrand entered his second cohabitation with his prestige damaged by his party's recent misfortunes. He had also lost stature by a mistaken judgment in his own "reserved" sector of foreign policy. He had been a leading drafter of the Maastricht Treaty (1991), designed to strengthen the institutional structures of the European Community. When the treaty encountered hostile criticism, he gambled on a popular referendum in France to bolster support. The outcome was a bare 51 percent approval by the French voters, and, although it was enough to put Maastricht into effect, the evidence of deep division in France further reduced the president's prestige. Still another embarrassment was the revelation in 1994 that Mitterrand had accepted a bureaucratic post in Pétain's Vichy regime in 1942–43. There were cries of outrage from Socialists and former resisters, yet the shock and fury quickly faded. In some circles he was credited with throwing his critics off balance by his clever management of the news. Prior to his death in January 1996, Mitterrand left his mark culturally on Paris as well, where grandiose architecture projects such as the Opéra de la Bastille, the expanded Louvre, the towering Grande Arche de la Défense, and the new Bibliothèque Nationale de France became his legacy.

Mitterrand's second cohabitation (1993–95) proved more helpful to Prime Minister Balladur than to the president. It also proved deeply disappointing to Jacques Chirac, who had engineered Balladur's appointment on the assumption that the latter would step aside when the presidential election approached. Chirac had failed to see that his stylish and courteous stand-in might develop into his own most serious rival. By 1995 Balladur was the clear front-runner and announced his presidential candidacy against his own party leader, Chirac. Meanwhile, the Socialists, after some initial scrambling to find a viable candidate, endcd by choosing party official Lionel Jospin, who led the field in the first round of voting. Chirac, a vigorous campaigner, outpaced Balladur, and in the runoff he won again, this time against Jospin. His victory brought to an end the 14-year Socialist presidency. (G.Wr.)

Chirac's presidency

The right-of-centre triumph did not last. In the anticipated elections that Chirac called in 1997, a Socialist majority swept back to power, and Jospin returned to head a coalition of Socialists, Communists, and Greens. Whereas the policies of Mitterrand's second term had made concessions to the free market, Chirac's moderate prime minister, Alain Juppé (1995–97), had made serious concessions to the welfare state. Under Jospin, as under Juppé, pragmatic cohabitation struggled to maintain both economic growth and the social safety net. Privatization proceeded apace, inflation remained under control, and the introduction of the euro (the single European currency) in January 1999 boosted competition and investment. Unemployment, however, stubbornly hovered around 12 percent in the last decade of the century, casting doubt on Jospin's hope that growth and social progress would be reconciled. In 2002 Chirac won reelection over extremist Le Pen, whose surprisingly strong showing in the first round of voting led Jospin to announce his resignation. This socioeconomic balancing act remained in place, though, pitting the popularity of progressive social legislation against the difficulties of high taxes, restrictive social security demands on employers, and precarious funding for health and welfare projects.

When France hosted and won the football (soccer) World Cup in 1998, it had been a triumph not only for national sporting pride but for cohabitation at the highest levels, as it showcased multiracial cooperation on a winning squad made up of Arabs, Africans, and Europeans, reflecting France's increasingly diverse society.

France took the world spotlight again in 2003, when the Chirac administration—believing the regime of Iraqi leader Saddam Hussein to be cooperating with United Nations inspectors searching for weapons of mass destruction—led several members of the UN Security Council in effectively blocking UN authorization of the use of force against Iraq.

Two years later, French triumph in diversity crumbled when the accidental deaths of two immigrant teenagers sparked violence in Paris that spread rapidly to other parts of the country in the weeks that followed. Nearly 9,000 cars were torched and nearly 3,000 arrests made during the autumn riots, which were fueled by high unemployment, discrimination, and lack of opportunity within the immigrant community.

Society since 1940. The surge of economic growth that lasted from the mid-1950s to the mid-1970s brought extensive changes in French lifestyles and in some of the society's basic structures. As the century neared its end, most French people had come to enjoy greater comfort and security than their forebears; they took for granted automobiles, modern household appliances, and vacation homes in the country, which had been regarded as luxuries not too long before. The French had been converted to installment buying and supermarket shopping; they spent lcss on food and drink and more on health and leisure. Thanks to the social security system that was expanded after World War II, they were better-protected against the hazards of illness, unemployment, and a neglected old age.

Urbanization

The most striking structural change taking place in France was rapid urbanization. The farm population, which stood at about one-third of the total population in 1940, fell to less than 5 percent in the 1990s; yet farm production increased as modern techniques spread, making France one of the world's leading agricultural exporters. In the industrial regions, modern technology and a new managerial spirit brought France to the threshold of the postindustrial age. The proportion of unskilled workers declined in favour of technically trained specialists, and even more dramatic was the explosive growth in the numbcr of white-collar employees and middle-lcvcl managcrs. At thc basc of this social pyramid was a new prolctariat of immigrants from southern Europe and Africa, who provided the manual labour that most French workers were no longcr willing to perform. In the 1990s thcse immigrants constituted between 5 and 10 percent of France's population, and their presence, aggravated by widespread joblessness, fed social and racial tensions. Anti-immigrant rcsentmcnt spurred the rise of Le Pen's National Front, which called for casting out aliens and reclaiming "France for the French" but benefited more from morosc protest against thc sitting government than from prcjudice. In 1999 the National Front, which always stood more for protcst than principle, succumbcd to internal dissensions and broke apart. With radical factions on the political far right and far lcft in disarray, the next test of French socicty, as of thc national economy, would come from a Europe henceforth without borders or national currencies—where workers, students, businesses, and immigrants from beyond the European Union could move freely from one country to another.

The cultural scene. Paris after World War II quickly regained its stature as one of the world's great centrcs of intellectual creativity. A cluster of brilliant thinkers and writers competed for influence, attracting acolytes both in France and abroad. The first postwar wave was lcd by Jean-Paul Sartre, whose influence made existentialism the leading ideology of the time. Sartre saw the world as "absurd" and irrational, lacking guideposts for humans adrift in a meaningless universe. People, said Sartre, know only that they exist and are free to cast their own lot. In the absence of any guiding power, individuals are condemned to freedom (hence responsibility), forced to forge their own lives, however insecure and contingent these may be, and to give them meaning by commitment to a course of action. Sartre's essays and novels made him the most admired intellectual of his generation and won him the Nobel Prize for Literature in 1964 (which he refused). His rival

Albert Camus, also a Nobel Prize winner, broke with Sartre over the latter's support of the Soviet Union and over Sartre's inability to define an ethical base for commitment to a cause. Camus's agnostic humanism led him to insist that even in an absurd world commitment must rest on clearly defined ethical principles—on the need to resist oppressors and fanatics and to respect the shared humanity of all people.

The dark postwar mood that lent existentialism its appeal faded when economic recovery set in. In the 1960s it was replaced by a new vogue called structuralism, whose scientific aspirations better suited a technological age. Drawing on the ideas of anthropologist Claude Lévi-Strauss, the structuralists stressed the persistence of "deep structures" that were held to underlie all human cultures through time, leaving little room for historical change or human initiative.

Structuralism

For a time structuralism became the dominant intellectual wave both in France and abroad; it showed signs of crystallizing into an ideology or worldview. But by the 1970s it gave way to a cluster of doctrines loosely labeled "post-structuralist," each variety identified with its own master-thinker: the philosopher Jacques Derrida, the intellectual historian Michel Foucault, the psychoanalyst Jacques Lacan, and the Marxologist Louis Althusser.

The structuralist vogue also affected the novelists who, beginning in the mid-1950s, launched *le nouveau roman*. More interested in theory and the subversive play of language than in storytelling, Alain Robbe-Grillet, Nathalie Sarraute, Michel Butor, and their imitators attracted much media and critical attention; but their provocative output stimulated more publicity than sales. Their iconoclastic aspirations were paralleled by those of a *nouvelle vague* of filmmakers such as Claude Chabrol, Jean-Luc Godard, Alain Resnais, and François Truffaut, whose movies of the late 1950s, '60s, and '70s revolutionized French cinema. "New novels" and "new wave" films may be compared to another contemporary creation popularized by the media: nouvelle cuisine, whose aesthetic objectives also evoked more critical than gourmandizing interest.

Those discouraged by pretentious fiction were turning to biography and general history, a realm dominated by the contributions of scholars such as Fernand Braudel, Emmanuel Le Roy Ladurie, and Pierre Nora. Marked by pathbreaking investigation of long-term perspectives and by a vivid, seductive style, their explorations of social, cultural, and economic history proved broadly appealing.

High culture has always seeped into popular culture and coloured it, perhaps more so in France than in other countries. Today, in France as elsewhere, the reverse is also true: the culture of everyday life encourages dislocations that elude socioeconomic and national boundaries. Differences of taste or of opinion, once dismissed as superficial, aggravate moral and political rifts. Sponsored by past Socialist governments as popular art, pop music in the 1960s called *yé-yé* ("yeah-yeah") and hip-hop music and graffiti art at the end of the 20th century were perceived by some as playful and by others as threatening. Multiculturalism was both welcomed as emancipating and scorned as divisive, as was the diffuse anti-Americanism, which for many stood in for antimodernism. All these disruptions were by-products of accelerated societal change, yet on all these cultural fronts French works and ideas continued to generate worldwide attention and, often, imitation.

(G.Wr./Eu.W.)

For later developments in the history of France, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 912, 921, 923, 924, 961, 963, and 972, and the *Index*.

BIBLIOGRAPHY

Geography. *General works:* Comprehensive surveys of the country, using both descriptive and analytic approaches, include PHILIPPE PINCHEMEL *et al., France: A Geographical, Social, and Economic Survey* (1987; originally published in French, 1964); CHRISTOPHER FLOCKTON and ELEONORE KOFMAN, *France* (1989); HILARY P.M. WINCHESTER, *Contemporary France* (1993); and XAVIER DE PLANHOL, *An Historical Geography of France* (1994; originally published in French, 1988), all with important bibliographies. See also MICHELIN PNEU, *Michelin: atlas routier et touristique*, 3rd ed. (1999), a useful reference source with more detailed information than its title suggests.

Land: Characteristics of major physical regions of France are provided in CLIFFORD EMBLETON (ed.), *Geomorphology of Europe* (1984); and GÉRARD MOTTET, *Géographie physique de la France*, 3rd rev. and enlarged ed. (1999). COMITÉ NATIONAL FRANÇAIS DE GÉOGRAPHIE, *Atlas de la France rurale: les campagnes françaises* (1984), contains maps and photographs illustrating aspects of rural France, both by type of agriculture and by region; and ANDRÉ BRUN *et al., Le Grand Atlas de la France rurale* (1989), provides documentation on all aspects of life from the point of view of agricultural geography. IAN SCARGILL, *Urban France* (1983), includes case studies of new towns.

Human geography: General overviews of regional planning, demography, social structure, economic conditions, culture, and politics are found in JOHN ARDAGH, *France in the New Century: Portrait of a Changing Society* (1999); J.E. FLOWER (ed.), *France Today: Introductory Studies*, 8th ed. (1997); ROBERT GILDEA, *France Since 1945* (1996, updated 2002); and EMMANUEL TODD, *La Nouvelle France* (1988, reissued 1990). Urban France is discussed in MARYSE FABRIÈS-VERFAILLIE, PIERRE STRAGIOTTI, and ANNIE JOUVE, *La France des villes: le temps des métropoles*, 2nd ed. (2000). Demographic analyses are provided in DANIEL NOIN and YVAN CHAUVIRÉ, *La Population de la France*, 5th ed., updated (1999); and ALEC G. HARGREAVES, *Immigration, 'Race,' and Ethnicity in Contemporary France* (1995).

Economy: Comprehensive surveys of modern economic conditions include JOHN TUPPEN, *The Economic Geography of France* (1983); JEAN-FRANÇOIS ECK, *La France dans la nouvelle économie mondiale*, 3rd ed. updated (1998); MARCEL BALESTE, *L'Économie française*, 13th ed. rev. (1995); and JOHN TUPPEN, *France Under Recession, 1981–1986* (1988). Additional surveys include COLIN GORDON and PAUL KINGSTON, *The Business Culture in France* (1996); JOSEPH SZARKA, *Business in France: An Introduction to the Economic and Social Context* (1992); DOMINIQUE TADDEI and BENJAMIN CORIAT, *Made in France: l'industrie française dans la competition mondiale* (1993); and JEAN JACQUES TUR, *Géographie humaine et économique de la France* (1994). Comprehensive regional studies can be found in MARCEL BALESTE *et al., La France: les 22 régions*, 5th ed. (2001); and PIERRE ESTIENNE, *Les Régions françaises*, 2 vol. (1999). Current developments are brought together in ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *OECD Economic Surveys: France* (annual). Studies of special features of the French economy include DOMINIQUE GAMBIER and MICHEL VERNIERES, *L'Emploi en France*, new ed. (1998); NACIMA BARON-YELLÈS, *Le Tourisme en France: territoires et stratégies* (1999); LAURENCE BANCEL-CHARENSOL, JEAN-CLAUDE DELAUNAY, and MURIEL JOUGLEUX, *Les Services dans l'economie française* (1999); and PIERRE BLOC-DURAFFOUR, *L'Industrie française* (1999).

Politics and society: Studies of political administration include SUDHIR HAZAREESINGH, *Political Traditions in Modern France* (1994); ALISTAIR COLE, *French Politics and Society* (1998); JAMES F. MCMILLAN, *Twentieth-Century France: Politics and Society 1898–1991* (1992); NICK HEWLETT, *Modern French Politics: Analysing Conflict and Consensus Since 1945* (1998); and MAURICE LARKIN, *France Since the Popular Front: Government and People 1936–1996*, 2nd ed. (1997). Law is treated in CHRISTIAN DADOMO and SUSAN FARRAN, *The French Legal System*, 2nd ed. (1996); and JOHN BELL *et al., Principles of French Law* (1998). The more specific question of decentralization is examined in JEAN MARC OHNET, *Histoire de la décentralisation française* (1996). Health care is discussed in MARC DURIEZ *et al., Le Système de santé en France*, 2nd updated ed. (1999). Social conditions and organizations in the social sphere are the focus of PIERRE LAROQUE (ed.), *Les Institutions sociales de la France*, updated ed. (1963, reissued 1980). A general view of contemporary features of French society is given in INSEE, *Données sociales: la société française* (1999), as well as *France, portrait social 1998–1999*, 2nd ed. (1999). A detailed discussion of urban policy is provided in ANTOINE ANDERSON, *Politiques de la ville* (1998).

Culture: Philosophical perspectives of French cultural life are examined in ALAIN FINKIELKRAUT, *The Defeat of the Mind* (1995; also published as *The Undoing of Thought*, 1988; originally published in French, 1987). BERNARD-HENRI LÉVY, *Éloge des intellectuels* (1987), is an introduction to intellectual life. DENIS HOLLIER (ed.), *A New History of French Literature* (1989, reissued 1994), discusses the literature from 842 onward. General views of the culture are given in JILL FORBES and MICHAEL KELLY (eds.), *French Cultural Studies: An Introduction* (1995); MAX SILVERMAN, *Facing Postmodernity: Contemporary French Thought on Culture and Society* (1999); MARTIN COOK (ed.), *French Culture Since 1945* (1993); and ALEX HUGHES and KEITH READER, *Encyclopedia of Contemporary French Culture* (1998, reissued 2002). (T.H.El./J.N.T./J.F.P.B./Jo.E.F.)

History. *General works:* A great historian's review of the long sweep of French history, from prehistoric times to the mod-

ern era, is FERNAND BRAUDEL, *The Identity of France*, 2nd ed. (1988–90; originally published in French, 1986), and *The Identity of France: People and Production* (1990; originally published in French, 1986). Excellent thematic treatments are provided in MARC BLOCH, *French Rural Society: An Essay on Its Basic Characteristics* (1966; originally published in French, 1931), a classic study; FERNAND BRAUDEL and ERNEST LABROUSSE (eds.), *Histoire économique et sociale de la France* (1970–82); GEORGES DUBY and ARMAND WALLON (eds.), *Histoire de la France rurale*, 4 vol. (1975–76, reprinted 1992); and GEORGES DUBY (ed.), *Histoire de la France urbaine*, 5 vol. (1980–85).

Gaul: A stimulating overview of Gaul in the context of French prehistory and early history is found in J.M. WALLACE-HADRILL and JOHN MCMANNERS (eds.), *France: Government and Society*, 2nd ed. (1970). Outstanding investigations of particular sites and areas of relevance to the study of Gaul as a whole are EDITH MARY WIGHTMAN, *Gallia Belgica* (1985); and A.L.F. RIVET, *Gallia Narbonensis: With a Chapter on Alpes Maritimae: Southern France in Roman Times* (1988). Life in later Roman Gaul is studied in JOHN MATTHEWS, *Western Aristocracies and Imperial Court, A.D. 364–425* (1975, reissued 1998); and RAYMOND VAN DAM, *Leadership and Community in Late Antique Gaul* (1985, reissued 1992). (J.F.Dr./J.D.P.)

Merovingian and Carolingian age (Early Middle Ages): A comprehensive introduction to the period is found in J.M. WALLACE-HADRILL, *The Barbarian West, 400–1000*, 3rd rev. ed. (1998). The Merovingians are considered in PATRICK J. GEARY, *Before France and Germany* (1988), and J.M. WALLACE-HADRILL, *The Long-Haired Kings and Other Studies in Frankish History* (1962, reprinted 1982). PIERRE RICHÉ, *The Carolingians: A Family Who Forged Europe* (1993; originally published in French, 1983) focuses on the Carolingians. Special studies of the civilization of the period include J.M. WALLACE-HADRILL, *The Frankish Church* (1983); SUZANNE FONAY WEMPLE, *Women in Frankish Society: Marriage and the Cloister, 500 to 900* (1981, reprinted 1985); and GEORGES DUBY, *The Early Growth of the European Economy: Warriors and Peasants from the Seventh to the Twelfth Century* (1974, reissued 1978; originally published in French, 1973). (G.Fo./B.S.Ba./J.D.P.)

The emergence of France, c. 850–1180: This period as a whole is well treated in JEAN DUNBABIN, *France in the Making, 843–1180* (1985). The political history of this and the succeeding periods is surveyed best in ELIZABETH M. HALLAM and JUDITH EVERARD, *Capetian France, 987–1328*, 2nd ed. (2001); but ROBERT FAWTIER, *The Capetian Kings in France: Monarchy & Nation, 987–1328* (1960, reissued 1982; originally published in French, 1942), is still a useful classic. Social change and feudalization are studied in MARC BLOCH, *Feudal Society* (1961, reprinted 1989; originally published in French, 1939–40), a seminal work. An alternative to Bloch's model has been developed by GEORGES DUBY, *La Société aux XI^{e} et XII^{e} siècles dans la région mâconnaise* (1953, reissued 1988), and *The Three Orders: Feudal Society Imagined* (1980; originally published in French, 1978). DOMINIQUE BARTHELEMY, *La Mutation de l'an mil, a-t-elle eu lieu?: servage et chevalerie dans la France des X^{e} et XI^{e} siècles* (1997).

France in the later Middle Ages, 1180–1490: The general political history of the period cited is covered in the works of HALLAM and EVERARD and FAWTIER listed in the previous section. Individual reigns are studied in JOHN W. BALDWIN, *The Government of Philip Augustus: Foundations of French Royal Power in the Middle Ages* (1986); and WILLIAM CHESTER JORDAN, *Louis IX and the Challenge of the Crusade: A Study in Rulership* (1979). Economy and society are studied in GEORGES DUBY, *Rural Economy and Country Life in the Medieval West* (1968, reprinted 1998; originally published in French, 1962). Social unrest is discussed in MICHEL MOLLAT and PHILIPPE WOLFF, *The Popular Revolutions of the Late Middle Ages* (1973; originally published in French, 1970). CHARLES PETIT-DUTAILLIS, *The French Communes in the Middle Ages* (1978; originally published in French, 1947), remains the standard account on the cities and towns. (T.N.B./J.D.P.)

France from 1490 to 1715: JANINE GARRISSON, *History of Sixteenth-Century France* (1995; originally published in French, 1991); and EMMANUEL LE ROY LADURIE, *The Royal French State, 1460–1610* (1994; originally published in French in 1987), offer clear summaries of French history from the Renaissance through the Wars of Religion. DAVID POTTER, *A History of Modern France, 1460–1560* (1995), is institutional rather than chronological in approach. J. RUSSELL MAJOR, *Representative Government in Early Modern France* (1980), and *From Renaissance Monarchy to Absolute Monarchy: French Kings, Nobles & Estates* (1994), explain the role of representative assemblies. SARAH HANLEY, *The Lit de Justice of the Kings of France: Constitutional Ideology in Legend, Ritual, and Discourse* (1983), analyzes one particularly important aspect of the relationship between monarch and institutions. See also R.J. KNECHT, *Renaissance Warrior and Patron: The Reign of Francis I*, rev. and expanded ed. (1994); and DAVID PARKER, *Class and State in Ancien Régime France: The Road to Modernity?* (1996). DONALD R. KELLEY, *The Beginning of Ideology: Consciousness and Society in the French Reformation* (1981), is a study of political thought.

ROBIN BRIGGS, *Early Modern France, 1560–1715*, 2nd ed. (1998), is a reliable introduction to 17th-century history. On the religious wars, MACK P. HOLT, *The French Wars of Religion, 1562–1629* (1995), is an able synthesis. BARBARA B. DIEFENDORF, *Beneath the Cross* (1991), shows the social origins of the rival religious parties. ROLAND MOUSNIER, *The Assassination of Henry IV* (1973; originally published in French, 1964), is a brilliant analysis of the intersection of politics and religion at the beginning of the 17th century.

YVES-MARIE BERCÉ, *The Birth of Absolutism: A History of France, 1598–1661* (1996; originally published in French, 1992), is an admirably clear narrative of the rise of the absolutist monarchy. JAMES B. COLLINS, *The State in Early Modern France* (1995, reissued 1999), is an overview of the development of monarchical institutions from 1600 to the Revolution. DAVID BUISSERET, *Henry IV* (1984, reissued 1992), is a good life of the monarch who ended the religious wars. A. LLOYD MOOTE, *Louis XIII, the Just* (1989), brings Henri IV's enigmatic successor to life. Other works on the period's political developments include JOSEPH BERGIN, *Cardinal Richelieu: Power and the Pursuit of Wealth* (1985); J.H. SHENNAN, *The Parlement of Paris*, rev. ed. (1998); N.M. SUTHERLAND, *The French Secretaries of State in the Age of Catherine de Medici* (1962, reprinted 1976); RICHARD BONNEY, *The King's Debts: Finance and Politics in France, 1589–1661* (1981), and *Society and Government in France Under Richelieu and Mazarin, 1624–61* (1988); and SHARON KETTERING, *Patrons, Brokers, and Clients in Seventeenth-Century France* (1986). YVES-MARIE BERCÉ, *History of Peasant Revolts* (1990; originally published in French, 1986), synthesizes scholarship on popular movements. OREST RANUM, *The Fronde: A French Revolution* (1993), is a concise summary of the confusion of conflicts from 1648 to 1653. GEOFFREY TREASURE, *Mazarin: The Crisis of Absolutism in France* (1995), is a thorough biography of the man who preserved Richelieu's handiwork and set the stage for Louis XIV's reign. ELIZABETH RAPLEY, *The Dévotes: Women and Church in Seventeenth-Century France* (1990, reissued 1993), highlights the role of women in the period's Catholic revival.

JAMES R. FARR, *Hands of Honor: Artisans and Their World in Dijon, 1550–1650* (1988); and GEORGES VIGARELLO, *Concepts of Cleanliness: Changing Attitudes in France Since the Middle Ages* (1988; originally published in French, 1985), discuss social conditions.

The Age of Louis XIV: PIERRE GOUBERT, *Louis XIV and Twenty Million Frenchmen* (1970; originally published in French, 1966), provides a synthesis of Louis's reign. FRANÇOIS BLUCHE, *Louis XIV* (1990; originally published in French, 1986), is the most recent full biography of France's most famous king. DANIEL DESSERT, *Louis XIV prend le pouvoir* (1989, reissued 2000), offers a revisionist account of the reign's beginnings. ALBERT N. HAMSCHER, *The Parlement of Paris After the Fronde, 1653–1673* (1976), and *The Conseil Privé and the Parlements in the Age of Louis XIV: A Study in French Absolutism* (1987), provide good introductions to the period. ROGER METTAM, *Power and Faction in Louis XIV's France* (1988); and ROGER METTAM (ed.), *Government and Society in Louis XIV's France* (1977), analyze the political structure. WILLIAM BEIK, *Absolutism and Society in Seventeenth-Century France* (1985), demonstrates in detail how Louis XIV's government enlisted the collaboration of the nobility. JOHN A. LYNN, *Giant of the Grand Siècle: The French Army 1610–1715* (1997), recounts the rise of the main tool of France's international power. WARREN C. SCOVILLE, *The Persecution of Huguenots and French Economic Development, 1680–1720* (1960), is important in assessing the effects of the revocation of the Edict of Nantes. PETER BURKE, *The Fabrication of Louis XIV* (1992), shows the interconnection between politics and art under the "Sun King." (J.H.Sh./J.D.P.)

France from 1715 to 1789: ALEXIS DE TOCQUEVILLE, *The Old Régime and the French Revolution* (1955, reprinted 1978; originally published in French, 1856), is still a basic source for the study of the period. Comprehensive histories include DANIEL ROCHE, *France in the Enlightenment* (1998; originally published in French, 1993), a detailed survey of the period's society and culture; and vol. 1 of ALFRED COBBAN, *A History of Modern France*, 3 vol. (1969). MICHEL ANTOINE, *Louis XV* (1989), is both a biography and an exhaustive account of high politics in that monarch's reign. GUY CHAUSSINAND-NOGARET, *The French Nobility in the Eighteenth Century* (1985; originally published in French, 1986), explains the changing nature of the country's ruling elite.

STEVEN LAURENCE KAPLAN, *Provisioning Paris: Merchants and Millers in the Grain and Flour Trade During the Eighteenth Century* (1984), describes a basic feature of the period's economy. Economic histories make much use of the 18th-century travelogue of ARTHUR YOUNG, *Travels During the Years 1787, 1788 & 1789: Undertaken More Particularly with a View of Ascertain-*

ing the Cultivation, Wealth, Resources, and National Prosperity of the Kingdom of France, 2nd ed., 2 vol. (1794), available in many later editions. There are fascinating glimpses of urban life in LOUIS-SÉBASTIEN MERCIER, *Panorama of Paris: Selections from Tableau de Paris* (1999; originally published in French, 1781); and JACQUES-LOUIS MÉNÉTRA, *Journal of My Life* (1986; originally published in French, 1982). The culture and ideology of the period are explored in ROBERT DARNTON, *The Forbidden Best-Sellers of Pre-Revolutionary France* (1995); LIESELOTTE STEINBRÜGGE, *The Moral Sex: Woman's Nature in the French Enlightenment* (1995; originally published in German, 1992); DENA GOODMAN, *The Republic of Letters: A Cultural History of the French Enlightenment* (1994); and THOMAS E. CROW, *Painters and Public Life in Eighteenth-Century Paris* (1985). DAVID BELL, *The Cult of the Nation* (2001), offers an important new analysis of the growth of national identity. A classic analysis of Rousseau's thought is JEAN STAROBINSKI, *Jean-Jacques Rousseau, Transparency and Obstruction* (1988; originally published in French, 1957). KEITH MICHAEL BAKER, *Inventing the French Revolution* (1990); and DALE K. VAN KLEY, *The Religious Origins of the French Revolution* (1996), show how revolutionary ideas developed out of prerevolutionary political discourse. (P.Hi./J.D.P.)

France from 1789 to 1815: Reliable overviews of the period include D.M.G. SUTHERLAND, *France 1789–1815: Revolution and Counter-Revolution* (1985); WILLIAM DOYLE, *The Oxford History of the French Revolution* (1989); and NORMAN HAMPSON, *A Social History of the French Revolution* (1963, reissued 1995).

The origins of the Revolution are considered in JEAN EGRET, *The French Pre-Revolution* (1977; originally published in French, 1962); and WILLIAM DOYLE, *Origins of the French Revolution*, 3rd ed. (1998). The best book on the Terror is still R.R. PALMER, *Twelve Who Ruled: The Year of the Terror in the French Revolution* (1941, reissued 1989). GEORGE RUDÉ, *Robespierre* (1967), provides excerpts from the Jacobin leader's speeches. MARTYN LYONS, *France Under the Directory* (1975), surveys the Revolution's later phase. FRANÇOIS FURET and MONA OZOUF (eds.), *A Critical Dictionary of the French Revolution* (1989; originally published in French, 1988), is an important and original collection of short essays on selected events, actors, institutions, ideas, and historians of the French Revolution. LYNN HUNT, *Politics, Culture, and Class in the French Revolution* (1984), analyzes the imagery and sociology of Revolutionary politics. Notable thematic studies include GEORGES LEFEBVRE, *The Great Fear of 1789: Rural Panic in Revolutionary France* (1973, reissued 1989; originally published in French, 1932); P.M. JONES, *The Peasantry in the French Revolution* (1988); ALBERT SOBOUL, *The Parisian Sans-culottes and the French Revolution, 1793–4*, trans. from French (1964, reprinted 1979); GEORGE RUDÉ, *The Crowd in the French Revolution* (1959, reprinted 1986); DOMINIQUE GODINEAU, *The Women of Paris and Their French Revolution* (1998; originally published in French, 1988); JOHN MCMANNERS, *The French Revolution and the Church* (1969, reprinted 1982); JEAN-PAUL BERTAUD, *The Army of the French Revolution* (1988; originally published in French, 1979); EMMET KENNEDY, *A Cultural History of the French Revolution* (1989); and JACQUES GODECHOT, *The Counter-Revolution: Doctrine and Action, 1789–1804* (1971, reissued 1981; originally published in French, 1961). The international dimension of the Revolution is interpreted in R.R. PALMER, *The World of the French Revolution* (1971). The best biography of a Revolutionary leader is LEO GERSHOY, *Bertrand Barère: A Reluctant Terrorist* (1962). JOHN HARDMAN, *Louis XVI* (1993), is a life of the movement's most illustrious victim; and the reasons for his fate are explained in DAVID JORDAN, *The King's Trial* (1979, reprinted 1993). ISSER WOLOCH, *The New Regime* (1994), shows how the Revolution's principles were institutionalized. JEREMY D. POPKIN, *Revolutionary News* (1990), explains the "media revolution" that was an integral part of the upheaval after 1789.

MARTYN LYONS, *Napoleon Bonaparte and the Legacy of the French Revolution* (1994), is a good overview. Napoleon's life is the subject of FELIX MARKHAM, *Napoleon* (1963); JEAN TULARD, *Napoleon: The Myth of the Saviour* (1984; originally published in French, 1977); and GEOFFREY ELLIS, *Napoleon* (1997, reissued 2000). The best volume on the Napoleonic regime in France is LOUIS BERGERON, *France Under Napoleon* (1981; originally published in French, 1972). OWEN CONNELLY, *Blundering to Glory: Napoleon's Military Campaigns*, rev. ed. (1999), is a critical and incisive analysis. (I.Wo./J.D.P.)

France since 1815: Vol. 2 and 3 of the already mentioned ALFRED COBBAN, *A History of Modern France*, 3 vol. (1957–62, reprinted 1969), present the period from the First Empire to the Republics in a sophisticated synthesis. GORDON WRIGHT, *France in Modern Times: From the Enlightenment to the Present*, 5th ed. (1995); and JEREMY D. POPKIN, *History of Modern France*, 2nd ed. (2001), are general surveys. Good introductions to the periods they cover are H.A.C. COLLINGHAM and R.S. ALEXANDER, *The July Monarchy 1830–1848* (1988); MAURICE AGULHON, *The Republican Experiment 1848–1852* (1983; originally published in French, 1973); ALAIN PLESSIS, *Rise and Fall of the Second Empire 1852–1871* (1985); JEAN-MARIE MAYEUR and MADELEINE RÉBERIOUX, *The Third Republic from Its Origins to the Great War 1871–1914* (1984; originally published in French, 1973); EUGEN WEBER, *The Hollow Years: France in the 1930s* (1994); JULIAN JACKSON, *France: The Dark Years, 1940–1944* (2002); JEAN-PIERRE RIOUX, *The Fourth Republic 1944–1958* (1987; originally published in French, 1980–83); and SERGE BERSTEIN, *The Republic of de Gaulle 1958–1969* (1993). SERGE BERSTEIN, *La France de l'expansion: l'apogée Pompidou 1969–1974* (1995), continues the story through the presidency of de Gaulle's successor. ANNIE MOULIN, *Peasantry and Society in France Since 1789* (1991; originally published in French, 1988); and GÉRARD NOIRIEL, *Workers in French Society in the 19th and 20th Centuries* (1990; originally published in French, 1986), are good introductions to social history.

Surveys of special topics on all or most of the period since 1815 include RENÉ RÉMOND, *The Right Wing in France from 1815 to de Gaulle*, 2nd ed. (1969; originally published in French, 1954), tracing change and continuity of the political right; GÉRARD CHOLVY and YVES-MARIE HILAIRE, *Histoire religieuse de la France contemporaine*, 3 vol. (1985–88, reissued 2000), an analysis of the role of various religions; and RAOUL GIRARDET, *La Société militaire dans la France contemporaine, 1815–1939*, updated ed. (1998), on the changing role and composition of the military corps. The role of France in world affairs is emphasized in PIERRE RENOUVIN, *Le XIX^e*, 2 vol. (1954–55), on the developments of the 19th century, part of the series *Histoire des relations internationales*. FRANÇOIS CARON, *An Economic History of Modern France*, trans. from French (1979, reissued 1983), revises older views about France's rate of growth. THEODORE ZELDIN, *France, 1848–1945*, 2 vol. (1973–77), explores modern French society, stressing its complexity and continuity. EUGEN WEBER, *Peasants into Frenchmen: The Modernization of Rural France: 1870–1914* (1976), argues that a sense of nationhood came to rural France only in the late 19th century.

Period studies include GUILLAUME DE BERTIER DE SAUVIGNY, *The Bourbon Restoration* (1966; originally published in French, 1955), the standard work on the period 1815–30; CLAIRE GOLDBERG MOSES, *French Feminism in the Nineteenth Century* (1984), which narrates the growth of women's movements; BONNIE G. SMITH, *Ladies of the Leisure Class* (1981), a model study of women's lives; DAVID H. PINKNEY, *Decisive Years in France, 1840–1847* (1986), arguing that France changed fundamentally in these years; PHILIP NORD, *The Republican Moment* (1995), which analyzes the rise of opposition to the Second Empire; MICHAEL B. MILLER, *The Bon Marché: Bourgeois Culture and the Department Store, 1869–1920* (1981); MICHAEL HOWARD, *The Franco-Prussian War: The German Invasion of France, 1870–1871*, 2nd ed. (2001), a model study; and EUGEN WEBER, *France fin de siècle* (1986). ROGER SHATTUCK, *The Banquet Years: The Origins of the Avant Garde in France, 1885 to World War I*, rev. ed. (1968, reissued 1984), is a brilliant survey of Parisian culture of the period. D.W. BROGAN, *France Under the Republic: The Development of Modern France (1870–1939)* (1940, reprinted 1974), is a classic account. JEAN-DENIS BREDIN, *The Affair: The Case of Alfred Dreyfus* (1986; originally published in French, 1983), provides a highly readable account of the great crisis.

The 20th century is studied in EUGEN WEBER, *Action Française: Royalism and Reaction in Twentieth Century France* (1962), a full analysis of this right-wing movement; ZEEV STERNHELL, *La Droite révolutionnaire, 1885–1914: les origines françaises du fascisme*, new ed., enlarged (2000), a controversial argument that fascism was born in France; JEAN-JACQUES BECKER, *The Great War and the French People* (1985, reissued 1993; originally published in French, 1980), which shows how France endured the ordeal of total war; ROBERT O. PAXTON, *Vichy France: Old Guard and New Order, 1940–1944* (1972, reissued 2001), a critical analysis of the Pétain regime; JEAN-BAPTISTE DUROSELLE, *La Décadence, 1932–1939*, 3rd rev. ed. (1985), and *L'Abîme: 1939–1945*, 2nd rev. ed. (1986, reissued 1990), 2 vol. of devastating analysis of French foreign policy before and during World War II; PHILIPPE BURRIN, *France Under the Germans* (1996; originally published in French, 1995), a synthesis of research on the occupation; CHARLES DE GAULLE, *War Memoirs*, 3 vol. (1955–60; originally published in French, 1954–60), and *Memoirs of Hope: Renewal and Endeavor* (1971; originally published in French, 2 vol., 1970–71), indispensable for an understanding of the Gaullist era; JEAN LACOUTURE, *Charles de Gaulle*, 3 vol. (1984–86), a full and perceptive biography; ROBERT GILDEA, *France Since 1945*, rev. ed. (2001), good on the postwar period; and ALFRED GROSSER, *Affaires extérieures: la politique de la France, 1944–1989* (1989), a penetrating analysis of postwar France's role in the world. JULIUS W. FRIEND, *The Long Presidency: France in the Mitterrand Years, 1981–1995* (1998), is a handy guide. HUGH DAUNCEY and GEOFF HARE (eds.), *France and the 1998 World Cup: The National Impact of a World Sporting Event* (1999), examines France as the host country and winner of the 1998 football (soccer) championships. (G.Wr./Eu.W.)

Franklin

Benjamin Franklin, next to George Washington possibly the most famous 18th-century American, by 1757 had made a small fortune, established the Poor Richard of his almanacs (written under the pseudonym Richard Saunders) as an oracle on how to get ahead in the world, and become widely known in European scientific circles for his reports of electrical experiments and theories. What is more, he was then just at the beginning of a long career as a politician, in the course of which he would be chief spokesman for the British colonies in their debates with the king's ministers about self-government and would have a hand in the writing of the Declaration of Independence, the securing of financial and military aid from France during the American Revolution, the negotiation of the treaty by which Great Britain recognized its former 13 colonies as a sovereign nation, and the framing of the Constitution, which for two centuries has been the fundamental law of the United States of America.

Franklin, portrait by Joseph-Siffred Duplessis, *c.* 1784.
By courtesy of the New York Historical Society

And as impressive as Franklin's public service was, it was perhaps less remarkable than his contributions to the comfort and safety of daily life. He invented a stove, still being manufactured, to give more warmth than open fireplaces; the lightning rod and bifocal eyeglasses also were his ideas. Grasping the fact that by united effort a community may have amenities which only the wealthy few can get for themselves, he helped establish institutions one now takes for granted: a fire company, a library, an insurance company, an academy, and a hospital. In some cases these foundations were the first of their kind in North America.

One might expect universal admiration for a man of such breadth and apparent altruism. Yet Franklin was disliked by some of his contemporaries and has ever since occasionally been attacked as a materialist or a hypocrite. D.H. Lawrence, the English novelist, regarded him as the embodiment of the worst traits of the American character. Max Weber, the German sociologist, made him the exemplar of the "Protestant ethic," a state of mind that contributed much, Weber thought, to the less admirable aspects of modern capitalism. Those who admire Franklin believe that his detractors have mistakenly identified him with Poor Richard, a persona of his own creation, or that they have relied too largely upon the incomplete self-portrait of his posthumously published *Autobiography.*

Youth in Boston

Early life (1706–23). Franklin was born in Boston, Massachusetts, on Jan. 17 (Jan. 6, Old Style), 1706, the 10th son of the 17 children of a man who was both soapmaker and candlemaker. He learned to read very early and had one year in grammar school and another under a private teacher, but his formal education ended when he was 10. At 12 he was apprenticed to his brother James, a printer. His mastery of the printer's trade, of which he was proud to the end of his life, was achieved between 1718 and 1723. In the same period he read tirelessly and taught himself to write effectively.

His first enthusiasm was for poetry, and in the first years of his apprenticeship he wrote two occasional ballads, no copies of which have survived. His father told him that "Verse-makers were always Beggars," and thereafter his interest in poetry was sporadic. Prose was another matter. *The Spectator,* Joseph Addison and Richard Steele's famous periodical of essays, had appeared in England in 1711–12 and was to be imitated for the greater part of a century but seldom with the persistence of Franklin, the printer's apprentice. He would read an essay, make a short note of the idea of each sentence, lay aside his notes for a few days, and then try to rewrite the essay. Comparison of his version with the original showed him the need to enlarge his vocabulary. Turning some *Spectator* papers into verse, and some days later reconverting them into prose, helped.

Influence of *The Spectator*

In 1721 James Franklin founded a *Spectator*-like weekly newspaper, the *New-England Courant,* to which readers were invited to contribute. Benjamin, now 16, read and perhaps set in type these contributions and decided that he could do as well himself. In 1722 he wrote a series of 14 essays signed "Silence Dogood." Satire of New England funeral elegies and of the lip service paid the learned languages at Harvard College foreshadowed later literary techniques to be used by Franklin.

Late in 1722 James Franklin got into trouble with the provincial authorities and was forbidden to print or publish the *Courant.* To keep the paper going, he discharged his younger brother from his original apprenticeship and made him the paper's nominal publisher. New indentures were drawn up but not made public. Some months later, after a bitter quarrel, Benjamin walked out, sure that James would not go to law and reveal the subterfuge he had devised. "It was not fair in me to take this Advantage," he wrote later, "and this I therefore reckon one of the first Errata [mistakes, in printer's lingo] of my Life."

Arrival in Philadelphia

Youthful adventures (1723–26). Failing to find work in Boston or New York City, Franklin proceeded to Philadelphia. One of the dramatic scenes of the *Autobiography* is the description of his arrival on a Sunday morning, tired and hungry. Finding a bakery, he asked for three pennies' worth of bread and got "three great Puffy Rolls." Carrying one under each arm and munching on the third, he walked up Market Street past the door of the Read family, where stood Deborah, his future wife. She saw him "& thought I made as I certainly did a most awkward ridiculous Appearance."

A few weeks later he was rooming at the Reads' and employed as a printer. By the spring of 1724 he was enjoying the companionship of other young men with a taste for reading and he was also being urged to set up in business for himself by the governor of Pennsylvania, Sir William Keith. At Keith's suggestion, Franklin returned to Boston to try to raise the necessary capital. His father thought him too young for such a venture, so Keith offered to foot the bill himself and arranged Franklin's passage to England so that he could choose his type and make connections with London stationers and booksellers. Franklin exchanged "some promises" with Deborah Read and, with a young friend, James Ralph, as companion, boarded the *London Hope* in November, expecting to find the letters of credit and introduction that Keith had promised. Not until the ship was well out at sea did he realize that the governor had not kept his promise. A fellow passenger, a Quaker merchant by the name of Thomas Denham, told him that Keith was unreliable; eventually Franklin could write charitably: "He wish'd to please every body; and, having little to give, he gave Expectations."

Years in London

In London Franklin quickly found employment in his trade and was able to lend money to Ralph, who was trying to establish himself as a writer. The two young men enjoyed the theatre and the other pleasures of the city; before long Ralph found a milliner for a mistress. When Ralph was in the country, teaching school, the milliner occasionally borrowed money from Franklin. "I grew fond of her Company," he remembered, "and being at this time under no Religious Restraints, & presuming on my Importance to her, I attempted Familiarities (another Erratum) which she repuls'd with a proper Resentment, and acquainted him with my Behaviour."

Still another "erratum" in retrospect was *A Dissertation on Liberty and Necessity, Pleasure and Pain* (1725), a deistical pamphlet he was inspired to write after having set type for William Wollaston's moral tract *The Religion of Nature Delineated.* Franklin argued therein that since man has no real freedom of choice he is not morally responsible for his actions, perhaps consoling himself for his treatment of Deborah, to whom he had written only once.

By 1726 Franklin was tiring of London. He considered becoming an itinerant teacher of swimming, but when Denham offered him a clerkship in his store in Philadelphia, with a prospect of fat commissions in the West Indian trade, he decided to return home.

Achievement of security and fame (1726–52). Denham died, however, a few months after Franklin entered his store. The young man, now 20, returned to his trade and in 1728 was able to set up a partnership with a friend. Two years later he borrowed money to become sole proprietor.

His private life at this time was extremely complicated. Deborah Read had married, but her husband had deserted her and disappeared. One matchmaking venture failed because Franklin wanted a settlement to pay off his business debt. A strong sexual drive, "that hard-to-be-govern'd Passion of Youth," was sending him to "low Women," and in the winter of 1730–31 he had a son, William, whose mother has never been identified. Franklin must have known that the child was expected when, his affection for Deborah having "revived," he "took her to Wife" on Sept. 1, 1730. Their common-law marriage lasted until Deborah's death in 1774. They had a son, who died at age four, and a daughter, Sarah, who survived them both. William was brought up in the household.

Wife and children

Franklin and his partner's first coup was securing the printing of Pennsylvania's paper currency. Franklin helped get this business by writing *A Modest Enquiry into the Nature and Necessity of a Paper Currency* (1729), and later he also became public printer of New Jersey, Delaware, and Maryland. Other money-making ventures included the *Pennsylvania Gazette,* published by Franklin from 1729 and generally acknowledged as among the best of the colonial newspapers, and the *Poor Richard's* almanacs, printed annually from 1732 to 1757. Some failures, of course, occurred: a German-language newspaper that lasted less than a year and a monthly magazine that expired after six issues in 1741. Franklin was nevertheless generally prosperous; he made enough to invest capital in real estate and in partnerships or working arrangements with printers in the Carolinas, New York, and the British West Indies. In 1748 he became a silent partner in the printing firm of Franklin and Hall, realizing in the next 18 years an average profit of almost £500 annually.

The first of his projects for social improvement by collective effort was the Junto, or Leather Apron club, organized in 1727 to debate questions of morals, politics, and natural philosophy and to exchange knowledge of business affairs. The need of Junto members for easier access to books led in 1731 to the organization of the Library Company of Philadelphia. Through the Junto, Franklin proposed a paid city watch, or police force. A paper read to the same group resulted in the organization of a volunteer fire company. In 1743 he called for a "constant correspondence" of men with scientific interests throughout the colonies, and later that year the American Philosophical Society was functioning. In 1749 he published *Proposals Relating to the Education of Youth in Pennsilvania;* in 1751, the Academy of Philadelphia, from which grew the University of Pennsylvania, was founded. So successful was Franklin as a promoter that anyone with a good cause in mind was likely to turn to him for help.

Franklin was also early involved in politics. He was clerk of the Pennsylvania legislature from 1736 until 1751 and postmaster of Philadelphia from 1737 until 1753. Prior to 1748, though, his most important political service was his part in organizing a militia for the defense of the colony against possible invasion by the French and the Spaniards, whose privateers were operating in the Delaware River. His skill in appealing to the self-interest of the various factions in the commonwealth is demonstrated in *Plain Truth; or, Serious Considerations on the Present State of the City of Philadelphia and Province of Pennsylvania* (1747).

Investigations in electrical science

In the 1740s electricity was a novel and fashionable subject. It was introduced to Philadelphians by an electrical machine sent to the Library Company by one of Franklin's English correspondents. In the winter of 1746–47, Franklin and three of his friends began to investigate electrical phenomena. The Philadelphia weather favoured them, as did the availability of talented instrument makers. Ingenious experiments and machines were devised and described in personal letters to England, which were relayed to the Royal Society of London or the *Gentleman's Magazine.* These papers were collected in 1751 as *Experiments and Observations on Electricity* and were translated into French (1752), German (1758), and Italian (1774).

Franklin's fame spread rapidly. The experiment he suggested to prove the identity of lightning and electricity was first made in France before he is believed to have tried the simpler but dangerous expedient of flying a kite in a thunderstorm. He and his associates concluded early that the "Electrical Fire" was "an Element diffused among, and attracted by other matter, particularly by Water and Metals." When a body with an overquantity approached one with an underquantity, a discharge equalized the electrical fire in the two. This "one fluid" theory accounted for more of the observable phenomena than had any previous hypothesis, and his suggestion that buildings be protected from lightning by erecting pointed iron rods proved both practical and dramatic. Franklin may not have been as original as some admirers have thought, and his collaborators may not have received their full share of credit, but he invented many terms still used in discussing electricity (positive, negative, battery, conductor, and so on) and described the experiments with lucidity.

Intercolonial concerns

Public service (1753–85). In 1753 Franklin became deputy postmaster general, in charge of mail in all the northern colonies. Thereafter he began to think in intercolonial terms. His "Plan of Union," adopted by the Albany Congress in 1754, would have established a general council, with representatives from the several colonies, to organize the common defense against the encroaching French and to supervise Indian relations with new settlements. Reason was on Franklin's side, but neither the colonial legislatures nor the king's advisers were ready for such union, and this conflict has been regarded by some authorities as the key to his entire political career.

In 1755 Franklin was nearly ruined when he promised to stand good for the loss of horses and wagons supplied by Pennsylvania farmers to support General Edward Braddock's ill-fated campaign against Fort-Duquesne in the French and Indian War. For more than two months he faced the possibility of having to pay almost £20,000 out of his own pocket. The government eventually paid.

The need for funds to defend the frontier led the Pennsylvania legislature to seek to tax the lands of the Penn family, the proprietors under the colony's charter. Either their consent or a change in the form of government was required. In the spring of 1757 Franklin was chosen to represent the legislature in this matter, which occupied him in London for most of his time until August 1762. He negotiated a compromise, under which the Penns agreed to taxation of improved lands but not those unsurveyed. During this first mission he made close friends in England

and wrote *The Interest of Great Britain Considered with Regard to Her Colonies and the Acquisitions of Canada and Guadaloupe* (1760). It was designed to urge the annexation of Canada when the war was over. There were Englishmen who preferred to leave Canada to the French, as a check on the growing strength of the 13 colonies. A simpler check, Franklin wrote, would be for Parliament to pass a law requiring midwives to stifle every third or fourth child as soon as it was born.

The Treaty of Paris (1763), ending the Seven Years' War, gave Canada to Great Britain. By that time Franklin was back in Philadelphia, where, in conflict with the proprietors, the legislature decided that Pennsylvania ought to become a crown colony, and by the end of 1764 Franklin was back in London to negotiate in vain for a new charter.

Role in the Stamp Act crisis

The tribulations of Pennsylvania were submerged, however, in the flood of feeling surrounding the so-called Stamp Act crisis. Franklin opposed the Stamp Act, asserting that taxation ought to be the prerogative of the representative legislatures, but when it had been passed he made the mistake of underestimating American emotions; he ordered stamps for Franklin and Hall and nominated a friend for the post of stamp officer in Philadelphia. His fellow citizens were so outraged that Deborah, fearful of her house being mobbed, called on male relatives for armed defense. In London Franklin, recognizing his error, quickly did an about-face and threw himself into the campaign for repeal of the statute. He regained his prestige by a dramatic appearance before the House of Commons, where he answered 174 questions from an audience partly friendly and partly hostile. The stenographic report of the exchange showed him returning often to the right of the colonies to levy internal taxes by their own legislation.

Although he failed to get the new charter, Franklin was kept on as London agent for Pennsylvania, and three other colonies relied on him to represent their interests—Georgia (1768), New Jersey (1769), and Massachusetts (1770). With this support and that of the British Whigs, the party of industrialists and dissenters in favour of parliamentary and philanthropic reform, he weathered the succession of crises ending with armed clashes at Lexington and Concord. He was gradually forced to the realization that there could be no reconciliation and that his dream of a British empire of self-governing nations would not come true. He did his best to present the American case to the British. Between 1765 and 1775 he published 126 newspaper articles on current controversies. At the end he was bitter, in such articles as "Rules by Which a Great Empire May Be Reduced to a Small One" and "An Edict by the King of Prussia," both first printed in 1773. Taken together, they are a capsule history of the long-drawn-out contest. In January 1774, because he had helped publish the letters of Thomas Hutchinson, governor of Massachusetts, to his British superiors, Franklin was dismissed from the post office. In March 1775, aware that there might be war, he left for Philadelphia. The day after his arrival he was a delegate to the Second Continental Congress, for which he served on committees for the organization of a postal system and for the drafting of the Declaration of Independence and on a commission that vainly attempted to bring Canada into the war as an ally.

Mission in Paris

In September 1776, the Congress agreed to send a commission to France to seek economic and military assistance. As one of three commissioners, Franklin arrived in Paris just before Christmas and was immediately engaged in secret negotiations with Charles Gravier, Count de Vergennes, minister of foreign affairs. Spies and informers infested his house, but Franklin was soon the hero of France, personifying the unsophisticated nobility of the New World, leading his people to freedom from the feudal past. His portrait was everywhere, on objets d'art from snuffboxes to chamber pots, his society sought after by diplomats, scientists, Freemasons, and fashionable ladies alike. The adulation was not without its ridiculous side, but Franklin, with his fur hat and spectacles, rose to the occasion with wit and social grace.

The sought-for treaties were signed in February 1778 after the British general John Burgoyne and 5,000 men surrendered at Saratoga, N.Y., and it was clear that the rebellion would not be crushed easily. Substantial loans were given to the revolutionists, and by the final victory at Yorktown in 1781, an estimated 12,000 soldiers and 32,000 sailors had left France to support General George Washington.

Despite these strong bonds, the peace was difficult. Spain had entered the war in 1779, hoping to recover Gibraltar, but, because of the conflict of interests in Florida and Louisiana, refused to recognize American independence. France had guaranteed that there would be no separate peace. Franklin worked with Vergennes until his fellow commissioners, John Adams and John Jay, overruled him on procedure, signing preliminary agreements with Great Britain late in 1782 without prior consultation with France. The formal treaty was signed Sept. 3, 1783.

Franklin wanted to return home but was kept in Paris for two more years to help make trade treaties. His popularity unabated, he observed the first balloon ascension and served on a committee appointed by Louis XVI to report on "animal magnetism," or hypnotism, thought by a German physician to cure many, if not all, diseases.

Role in the U.S. Constitutional Convention

Last years (1785–90). At 79, with a large stone in his bladder that made travel by carriage an agony, Franklin was carried to the port of Le Havre in a litter. Back in Philadelphia he lived quietly but continued to take some part in public life. His most important service was as a member of the Constitutional Convention of 1787. There he failed to convince his associates that an executive committee would be better than a president as head of state and that there should be a unicameral legislature. On the last day of the convention, a colleague read for him a plea that objections to the new form of government, his own among them, should be forgotten and that delegates should unanimously support the instrument that they had hammered out. Franklin's motion was promptly carried.

For the last year of his life he was bedridden, escaping severe pain only by the use of opium. He died on April 17, 1790, aged 84. Philadelphia gave him the most impressive funeral the city had ever seen, and in France, where Louis XVI was imprisoned, eulogies poured forth to the man who, to the French, was the symbol of enlightenment and freedom. All Europeans remembered the epigram of Anne-Robert-Jacques Turgot, the French economist: "He snatched the lightning from the skies and the sceptre from tyrants."

MAJOR WORKS

POLITICAL AND ECONOMIC: *A Modest Enquiry into the Nature and Necessity of a Paper Currency* (1729); *Plain Truth; or Serious Considerations on the Present State of the City of Philadelphia* (1747); *Proposals Relating to the Education of Youth in Pensilvania* (1749); *Observations Concerning the Increase of Mankind* (1755); *The Way to Wealth* (1757); *The Interest of Great Britain Considered with Regard to Her Colonies and the Acquisition of Canada and Guadaloupe* (1760); *Positions to be Examined Concerning National Wealth* (1769); *Journal of the Negotiations for Peace* (1782).

RELIGIOUS, PHILOSOPHICAL, AND SCIFNTIFIC: *A Dissertation on Liberty and Necessity, Pleasure and Pain* (1725); *Articles of Belief and Acts of Religion* (1728); *Experiments and Observations on Electricity* (1751).

OTHER WORKS: *Poor Richard's* (1732–57), an almanac that has many famous maxims; Franklin's *Autobiography* (1771–88); *Information to Those Who Would Remove to America* (1784).

BIBLIOGRAPHY. *The Papers of Benjamin Franklin*, 22 vol., ed. by LEONARD W. LABAREE *et al.* (1959–82), with additional volumes expected, is the definitive collection. *The Writings of Benjamin Franklin*, 10 vol., ed. by ALBERT H. SMYTH (1905–07, reprinted 1970), has heretofore been the chief collection. A full biography is CARL VAN DOREN, *Benjamin Franklin* (1938, reissued 1980); a good brief biography is VERNER W. CRANE, *Benjamin Franklin and a Rising People* (1954). See also RONALD W. CLARK, *Benjamin Franklin* (1983), a popular biography; THOMAS J. FLEMING, *The Man Who Dared the Lightning: A New Look at Benjamin Franklin* (1971); ARTHUR B. TOURTELLOT, *Benjamin Franklin: The Shaping of Genius, 1706–1723* (1977), a study of his heritage and youth; CLAUDE-ANNE LOPEZ and EUGENIA W. HERBERT, *The Private Franklin: The Man and His Family* (1975); BRIAN M. BARBOUR (ed.), *Benjamin Franklin: A Collection of Critical Essays* (1979), with emphasis on his roles as writer and shaper of the American national character; and BRUCE INGHAM GRANGER, *Benjamin Franklin, An American Man of Letters* (1964, new ed. 1976). (T.Hor./Ed.)

Frederick the Great

Frederick II the Great, third king of Prussia from 1740 to 1786, ranks among the two or three dominant figures in the history of modern Germany. Under his leadership Prussia became one of the great states of Europe. Its territories were greatly increased and its military strength displayed to striking effect. From early in his reign Frederick achieved a high reputation as a military commander, and the Prussian army rapidly became a model admired and imitated in many other states. He also emerged quickly as a leading exponent of the ideas of enlightened government, which were then becoming influential throughout much of Europe; indeed, his example did much to spread and strengthen those ideas. Notably, his insistence on the primacy of state over personal or dynastic interests and his religious toleration widely affected the dominant intellectual currents of the age. Even more than his younger contemporaries, Catherine II the Great of Russia and Joseph II in the Habsburg territories, it was Frederick who, during the mid-18th century, established in the minds of educated Europeans a notion of what "enlightened despotism" should be. His actual achievements, however, were sometimes less than they appeared on the surface; indeed, his inevitable reliance on the landowning officer (Junker) class set severe limits in several respects to what he could even attempt. Nevertheless, his reign saw a revolutionary change in the importance and prestige of Prussia, which was to have profound implications for much of the subsequent history of Europe.

By courtesy of the Staatliche Museen zu Berlin

Frederick II, portrait by Antoine Pesne (1683–1757). In the Gemäldegalerie, Berlin.

Tension between father and son

Early life. Frederick was born on Jan. 24, 1712, in Potsdam, near Berlin. He was the eldest surviving son of Frederick William I, king of Prussia, and Sophia Dorothea of Hanover, daughter of George I of Britain. Frederick's upbringing and education were strictly controlled by his father, who was a martinet as well as a paranoiac. Encouraged and supported by his mother and his sister Wilhelmina, Frederick soon came into bitter conflict with his father. Frederick William I deeply despised the artistic and intellectual tastes of his son and was infuriated by Frederick's lack of sympathy with his own rigidly puritanical and militaristic outlook. His disappointment and contempt took the form of bitter public criticism and even outright physical violence, and Frederick, beaten and humiliated by his father, often over trifling details of behaviour, took refuge in evasion and deceit. This personal and family feud culminated spectacularly in 1730, when Frederick was imprisoned in the fortress of Küstrin after planning unsuccessfully to flee initially to France or Holland. Lieutenant Hans Hermann von Katte, the young officer who had been his accomplice in the plan, was executed in Frederick's presence, and there was for a short time a real possibility that the prince might share his fate. During the next year or more Frederick, as a punishment, was employed as a junior official in local administration and deprived of his military rank. The effects of this terrible early life are impossible to measure with accuracy, but there is little doubt that the violent and capricious bullying of his father influenced him deeply.

In 1733, after a partial reconciliation with his father, Frederick was married to a member of a minor German princely family, Elizabeth Christine of Brunswick-Bevern, for whom he never cared and whom he systematically neglected. In the following year he saw active military service for the first time under the great Austrian commander Eugene of Savoy against the French army in the Rhineland. In the later 1730s, in semiretirement in the castle of Rheinsberg near Berlin and able for the first time to give free rein to his own tastes, he read voraciously, absorbing the ideas on government and international relations that were to guide him throughout his life. These years were perhaps the happiest that Frederick ever experienced. However, his relations with his father, though somewhat improved, remained strained.

Accession to the throne and foreign policy. Frederick William I died on May 31, 1740, and Frederick, on his accession, immediately made it clear to his ministers that he alone would decide policy. Within a few months he was given a chance to do so in a way that revolutionized Prussia's international position. The Holy Roman emperor Charles VI, of the Austrian house of Habsburg, died on October 20, leaving as his heir a daughter, the archduchess Maria Theresa, whose claims to several of the heterogeneous Habsburg territories were certain to be disputed. Moreover, her army was in a poor state, the financial position of the Habsburg government very difficult, and her ministers mediocre and in many cases old. Frederick, however, thanks to his father, had a fine army and ample funds at his disposal. He therefore decided shortly after the emperor's death to attack the Habsburg province of Silesia, a wealthy and strategically important area to which the Hohenzollerns, the ruling family of Prussia, had dynastic claims, though weak ones. The most important threat to his plans was Russian support for Maria Theresa, which he hoped to avert by judicious bribery in St. Petersburg and by exploiting the confusion that was likely to follow the imminent death of the empress Anna. He also hoped that Maria Theresa would cede most of Silesia in return for a promise of Prussian support against her other enemies, but her refusal to do so made war inevitable.

Invasion of Silesia

The first military victory of Frederick's reign was the battle of Mollwitz (April 1741), though it owed nothing to his own leadership; in October Maria Theresa, now threatened by a hostile coalition of France, Spain, and Bavaria, had to agree to the Convention of Klein-Schnellendorf, by which Frederick was allowed to occupy the whole of Lower Silesia. However, the Habsburg successes against the French and Bavarians that followed so alarmed Frederick that early in 1742 he invaded Moravia, the region south of Silesia, which was under Austrian rule. His rather incomplete victory at Chotusitz in May nonetheless forced Maria Theresa to cede almost all of Silesia by the Treaty of Berlin of 1742 in July. This once more allowed Habsburg forces to be concentrated against France and Bavaria, and 1743 and the early months of 1744 saw Maria Theresa's position in Germany become markedly stronger. Frederick, again alarmed by this, invaded Bohemia in August 1744 and rapidly overran it. However, by the end of the year lack of French support and threats to his lines of communication had forced him to retreat. Moreover, the elector Augustus III (king of Poland and the elector of Saxony) now joined Maria Theresa in attacking him in Silesia. He was rescued from this threatening situation by the prowess of his army; victories at Hohenfriedberg in June 1745 and at Soor in September were followed by a Prussian invasion of Saxony. The Treaty of Dresden, signed on Dec. 25, 1745, finally established Prussian rule in Silesia and ended for the time being the complex series of struggles that had begun five years earlier.

Silesia was a valuable acquisition, being more developed economically than any other major part of the Hohenzollern dominions. Moreover, military victory had now made Prussia at least a semigreat power and marked Frederick as the most successful ruler in Europe. He was well aware, however, that his situation was far from secure. Maria Theresa was determined to recover Silesia, and the peace she signed with France and Spain at Aix-la-Chapelle in 1748 allowed her to accelerate significant improvements in the administration of her territories and the organization of her army. Frederick's alliance with France, which dated from an agreement of June 1741, was based merely on mutual hostility toward the Habsburgs and had never been effective. More serious, anti-Prussian feeling was now running high in Russia, where both the empress Elizabeth, who had ascended the throne in 1741, and her chancellor, Aleksey Bestuzhev-Ryumin, bitterly disliked Frederick. Moreover, Great Britain, under George II, seeking an effective continental ally against France, seemed to be moving closer to Maria Theresa and Elizabeth. In September 1755 Britain signed an agreement with Russia by which Russia, in return for British subsidies, was to provide a large military force in its Baltic provinces to protect, if necessary, the electorate of Hanover, ruled by George II, against possible French or Prussian attack. Frederick was deeply alarmed by this: a hostile Austro-Russian alliance backed by British money seemed to threaten the destruction of Prussia. In January 1756 he attempted to escape from this menacing situation by an agreement with Britain for the neutralization of Germany in the Anglo-French colonial and naval war that had just begun. This, however, deeply antagonized Louis XV and the French government, who saw the agreement as an insulting desertion of France, Frederick's ostensible ally. The result was the signature in May of a Franco-Austrian defensive alliance. This did not in itself threaten Frederick, but he soon became convinced that a Russo-Austrian attack on him, with French support, was imminent. He determined to forestall his enemies and, in a daring move, invaded Saxony in August 1756 and marched on into Bohemia. This action has been more actively debated by historians than any other event of Frederick's reign because it raised in an acute form the general issue regarding the morality of preventive military action. Though Frederick took the offensive and thus unleashed a great military struggle, there is no doubt that he was by 1756 seriously threatened, indeed, even more seriously than he himself realized, and that his enemies, most of all the empress Elizabeth, meant to destroy Prussia's newly won international status.

Invasion of Saxony and Bohemia

The Seven Years' War, on which he embarked thus soon became a life-and-death struggle. In 1757 France, Sweden, Russia, and many of the smaller German states joined the ranks of his opponents, while the Prussian invasion of Bohemia collapsed after a serious defeat at Kolín in June. Brilliant victories over the French and Austrian armies, respectively, at Rossbach and Leuthen in November and December partially reestablished Frederick's position, but it still remained extremely precarious. Ruthless exploitation of every available resource (notably of much of Saxony, which was under Prussian military occupation during most of the war), debasement of the currency, and a British subsidy that he received in 1758–62 allowed Frederick with increasing difficulty to keep up the unequal struggle. More than anything, however, he was helped by the complete failure of his enemies to cooperate effectively, while a partly British and British-financed army in western Germany from 1758 onward neutralized the French military effort. Nevertheless, the strain was immense; in October 1757 a cabinet order suspended all payment of salaries and pensions to Prussian civil servants and judges apart from diplomats serving abroad. Frederick could still win victories in the field, as, for example, at Zorndorf (August 1758) against the Russians at heavy cost or at Liegnitz and Torgau (August and November 1760) against the Austrians. But he also suffered serious defeats at Hochkirch in October 1758 and above all at the hands of a Russian army at Kunersdorf in August 1759. This disaster temporarily reduced him to despair and thoughts of suicide; if it had been effectively followed up by his adversaries, he could not have continued the struggle. As the forces he could put in the field dwindled and resistance grew among his subjects to the unprecedented burdens imposed by the war (in 1760 the landowners of Brandenburg refused to contribute further), the Prussian position became increasingly difficult; by 1761 it was desperate. However, the death in January 1762 of the empress Elizabeth, the most bitter of all Frederick's enemies, completely changed the situation. Her successor, Peter III, a fanatical admirer of Prussia and Frederick, signed an armistice in May, followed by a Russo-Prussian peace treaty. This turn of events ended Maria Theresa's hopes of recovering Silesia. The Treaty of Hubertusburg (Feb. 15, 1763), which ended the war in Germany, left the province in Frederick's hands. Prussia had survived, and its military reputation was now greater than ever. The cost had been enormous, however. The Prussian army had lost 180,000 men during the struggle, and some Prussian provinces had been completely devastated.

The Treaty of Hubertusburg

Henceforth Frederick was determined to avoid another such conflict: the alliance with Russia that he signed in 1764 and which lasted until 1780 was directed largely to this end. Nevertheless, he still firmly opposed any growth of Habsburg power in Germany, and in July 1778 a new Austro-Prussian struggle broke out over the efforts of the emperor Joseph II, the son of Maria Theresa, to gain a large part of Bavaria. This War of the Bavarian Succession was half-hearted and short-lived, and the Treaty of Teschen ending it in May 1779 was a severe check to Joseph's ambitions and a diplomatic victory for Frederick. But this new conflict showed unmistakably that Austro-Prussian rivalry stemming from the events of 1740–41 was now a deeply ingrained fact of German political life. Fear of Habsburg ambitions continued to haunt Frederick to the end of his reign. His last significant achievement was to inspire the formation, in July 1785, of the League of Princes (Fürstenbund), which united a number of German states—the most important being Hanover, Saxony, and the archbishopric of Mainz—in successful opposition to Joseph II and his renewed efforts to acquire the whole of Bavaria in exchange for the Austrian Netherlands.

The first partition of Poland

The most important foreign policy development in the second half of Frederick's reign was the first partition of Poland, in 1772. By this Prussia gained the Polish province of West Prussia (though without the great commercial city of Danzig), and thus Brandenburg and Pomerania, the core of the monarchy, became linked with the theretofore isolated East Prussia. This gave the state a much greater territorial coherence and more defensible frontiers. It also moved its geographic centre decisively to the east and sharpened the social and political differences that tended to separate it from the states of western Europe.

Frederick had always hoped for territorial gains of this kind, and, as the weakness and confusion of the internally divided Polish republic increased during the 1760s, the possibilities of realizing them grew. In 1769 he tried indirectly to interest Catherine II of Russia in a partition but in vain. By January 1771, however, faced by strong Austrian opposition to her expansionist ambitions in southeastern Europe, the empress had changed her mind. The visit to St. Petersburg in that month of Frederick's younger brother Prince Henry played a decisive role in making a partition possible; the Habsburg government, which had hoped to recover Silesia or gain territory in the Balkans, was persuaded to join in the process. Frederick bears much of the responsibility for the partition, for he alone of the monarchs who took part had consciously desired it. Since both Russia and Austria were persuaded to follow a policy that was largely Prussian in inspiration, it ranks as perhaps his greatest diplomatic success.

Domestic policies. In administrative, economic, and social policy Frederick's attitudes were essentially conservative. Much of what he did in these areas was little more than a development of policies pursued by his father. He justified these policies in terms of the rationalizing rhetoric of "enlightened despotism," whereas the devoutly Protestant Frederick William I had done so in terms of religious obligation, but many of the objectives, and the means used to attain them, were the same. Frederick, in spite of his appalling personal relationship with his father, admired him as a ruler and freely acknowledged the debt he owed him. "Only his care," he wrote during the Seven Years' War, "his untiring work, his scrupulously just policies, his great and admirable thriftiness and the strict discipline he introduced into the army which he himself had created, made possible the achievements I have so far accomplished."

Concept of rulership

Like Frederick William I, Frederick thought of kingship as a duty. To him it entailed obligations to be met only by untiring and conscientious work. It was his duty to protect his subjects from foreign attack, to make them prosperous, to give them efficient and honest administration, and to provide them with laws that were simple and adapted to their wants and their particular temperament. In order to achieve these objectives, the ruler must sacrifice his own interests and any purely personal or family feeling. *Raison d'état,* the needs of the state, took precedence over these and also over the immediate comfort and happiness of his subjects. The ruler could carry out his duties effectively only if he kept the reins of government firmly in his own hands. His rule must be personal. He must not rely on ministers who were likely to be influenced by selfish ambitions or factional feeling and who might well keep important information from their master if they were allowed to. Personal rule alone could produce the unity and consistency essential to any successful policy. In his *Anti-Machiavel,* a somewhat conventional discussion of the principles of good government published in 1740 just before his accession, Frederick wrote that there were two sorts of princes—those who ruled in person and those who merely relied on subordinates. The former were "like the soul of a state" and "the weight of their government falls on themselves alone, like the world on the back of Atlas," whereas the second group were mere phantoms. Yet he would have rejected outright, and on the whole with justification, any suggestion that he ruled as a despot. On the contrary, he would have claimed that his power, however great, was exercised only within limits set by law and that the obligations inherent in his position made it impossible for him to govern in an arbitrary way.

The insistence that any effective monarchical rule must be intensely personal had obvious potential dangers. As Frederick grew older, these showed themselves with increasing clarity. His whole psychology was hostile to the development in the Prussian administration or army of any real originality, new ideas, or willingness to take initiatives or accept individual responsibilities. He fostered among those who served him a tendency to play safe and to perform their duties conscientiously but to do no more than that. Under him the Prussian administration was the most honest and hardworking in Europe. Its achievements, however, stemmed from the impetus supplied from above by the king rather than from any creative force inherent in the system itself. The provincial War and Domains Chambers established by Frederick William I in 1722 remained very important, and their number grew from 9 to 12. The General Directory, again created by Frederick William, as the main organ of central government with wide-ranging powers, acquired under Frederick several new departments (for commerce and manufactures in 1740, for mines and metallurgy in 1768, for forestry a few years later) but tended, as the reign went on, to become ossified and to lose a good deal of its former importance. The administration of Silesia after its acquisition in the 1740s was notably efficient, and its resources helped greatly in carrying Frederick through the dark days of the Seven Years' War. But tradition and continuity rather than innovation were the hallmarks of the Prussian administration under him; many of what new departures there were (for example, an effort in 1770 to introduce a system of state examinations for entry into the civil service) were not very effective. Many of the truly successful innovations were in the judicial system, where the reforming efforts of Samuel von Cocceji resulted in all judges in higher and appellate courts being appointed only after they had passed a rigorous examination. Cocceji also inspired the establishment in 1750 of a new Superior Consistory to supervise church and educational affairs and began the process of legal codification that culminated after Frederick's death in the issue of the Prussian Common Law (Das Allgemeine Preussische Landrecht) of 1794, one of the most important 18th-century efforts of this kind. Yet Frederick's unwillingness ever to admit a mistake or change his mind tended, as he grew older, to make the processes of government increasingly rigid and inflexible. The government's refusal to adapt and adjust, which was already visible during the monarch's later years, culminated in the Prussian collapse of 1806 before the armies of Napoleon.

The Prussian army

The overriding objective of Frederick's rule was to increase the power of the state. His desire to foster education and cultural life was sincere, but these humanitarian goals were secondary compared with the task of building a great army and gaining the financial resources needed to maintain it. The army was the pivot around which all else turned, and the administrative system existed essentially to recruit, feed, equip, and pay it. In proportion to the resources available to support it, its size was unequaled anywhere in Europe. In 1740 Frederick inherited a standing army of 83,000 men; when he died, this figure had risen to 190,000 (though of these only about 80,000 were Prussian subjects). Under him it remained a force of peasants and of numerous foreign recruits obtained often by outright kidnapping, officered by landowners. In Prussia the army was recruited almost entirely in the countryside; the function of townsmen was to pay for it through their taxes, not to serve in it. Up to a point Frederick tried to protect the peasants and the soldiers against the demands of the Junker landlord-officers. In 1749 and 1764 he issued decrees limiting the obligations of the peasant to his lord, and in 1748 he ordered officers not to treat their men "like serfs"; but these were essentially efforts to prevent the plight of the peasant from becoming so desperate that he would be driven into flight and thus jeopardize the supply of recruits. Throughout Frederick's reign, army service was for the majority of his subjects the most onerous of all the burdens imposed by the state. In order to finance the great army, heavy demands were made on territories that for the most part were poor. Nothing, however, seemed more important to the monarch than amassing a large reserve of cash to be used for the recruitment of men in case of war. The financial demands that a serious conflict would make were constantly on his mind, and the desperate struggles of 1756–62 confirmed him in his beliefs.

Much of the tax system, based on the excise (largely a tax on food) paid by the towns and the contribution (a complex property tax) raised in the countryside, supplemented by the profits of the extensive royal domains, remained essentially unchanged. Still, Frederick experimented with a number of new taxes, notably with a new system of taxing tobacco and some less important commodities (introduced in 1766 under the supervision of a French entrepreneur, Le Haye de Launay), but these innovations did not bring about significant changes. Indeed, many of Frederick's fiscal policies were ill-judged; for example, the maintenance of a great reserve of cash, which removed from circulation much of the liquid capital of a poor society, was economically damaging. Yet strict control of expenditure and relatively efficient tax collection meant that the government, unlike many others of the age, was never hamstrung by lack of money.

Frederick's economic policies were squarely in the mercantilist tradition. "The foundation of trade and manufactures," he wrote in his *Testament Politique* of 1752, "is to prevent money leaving the country and to make it come in." The direct and simplistic way in which these ideas were sometimes applied can be seen in an order of 1747 forbidding individuals to take more than 300 thalers

Economic policies

in specie out of their territories. So far as possible Prussia was to avoid importing foreign manufactured goods, and to this end domestic producers were to be helped by privileges and even outright grants of money. Exports were to be encouraged in the same way. In particular, much money was spent on efforts to develop a substantial silk industry, with generally disappointing results. By the end of the reign textiles of all kinds accounted for two-thirds of Prussia's industrial production, and the textile industry employed about 90 percent of the industrial labour force, but this situation owed little to Frederick's economic policies. Efforts to foster the production of porcelain—which, like silk, was one of the industrial status symbols of a number of 18th-century rulers—were also costly and not very effective. A small number of favoured industrialists, notably David Splitgerber and Johann Ernst Gotzkowsky in the 1750s, benefited by these policies, but for Prussia as a whole they were largely a misuse of resources. Other new creations such as the Maritime Trading Company (Seehandlung), a government-backed corporation set up in 1772 to develop overseas trade, and even the Royal Bank of Berlin, established in 1765, were also marginal to the economic life of Frederick's territories, which, except to some extent in Silesia, in the area around Berlin, and in the little county of Mark in western Germany, continued to be based on agriculture.

Some of the state's programs, however, achieved real success, though sometimes at high cost. Most important was the sustained effort, in the 1760s and '70s, to attract immigrants and to settle them on waste or depopulated land; this settlement program formed the central feature of the *rétablissement,* the making good of the losses of the Seven Years' War. During Frederick's reign more than 300,000 settlers were attracted to Prussia from other parts of Europe—a substantial addition to a population that in 1740 had numbered only about 2,200,000. In addition, the army provided a large market for arms and woolen cloth for uniforms and thus did something to stimulate economic growth. Moreover, in peacetime the soldiers served with their regiments only for a few months of the year, spending the remaining part in agriculture or some urban employment. The fact that they were in this way integrated into society helped to offset the burden that so great a military effort placed on the economy.

The importance of the nobility

Frederick's social policies were as conservative as his economic ones. He considered the nobility the most important class in Prussian society. From it were drawn the majority of the army officers and virtually all the higher-ranking ones. It also produced the majority of his officials and all his ministers and completely dominated local government in the countryside. In Frederick's eyes, the nobility alone of all the social groups had a sense of personal honour and responsibility. The continued existence of the state depended on it, and the regime could not function without its cooperation. Thus its interests were always to be safeguarded. In particular, it was not to be diluted by the grant of noble status to self-made bourgeois, and land owned by noble families was to be protected against purchase by members of the urban middle class, however wealthy. Frederick stated these ideas repeatedly in his voluminous writings on statecraft, notably in the political testaments of 1752 and 1768 drawn up for his successor. Given this attitude, it is not surprising that his reign saw little practical improvement for the peasantry, much of which, in Pomerania, Brandenburg, and East Prussia, was still personally unfree, owing labour services to noble landowners. In principle, Frederick sincerely disliked serfdom. In practice, however, he realized that any rapid move against it risked the disruption of Prussia's agricultural life and the erosion of the position of the all-important nobility. His efforts to improve the lot of his peasant subjects were therefore little more than gestures. As part of his commitment to stimulate recovery from the losses of the Seven Years' War, he tried to abolish serfdom in Prussian Pomerania and also to give the peasantry of Upper Silesia greater security of tenure, but none of this had much practical effect because he never contemplated any significant change in the social order.

Frederick prided himself on being, among rulers, the leading representative of the high culture of his day. He was a prolific writer on contemporary history and politics; his *Histoire de mon temps* (1746) is still a source of some value for the period it covers. He produced large quantities of mediocre poetry and composed music. He invited to Prussia several of the leading French intellectuals of the age, notably Voltaire (with whom he soon quarreled). But here again his outlook was essentially conservative. Culture to him meant French culture: he wrote and spoke French by preference, using German only when necessary. He had no interest in the profound intellectual stirrings occurring in Germany. Berlin under him never became an important intellectual centre. Gotthold Ephraim Lessing, perhaps the greatest German writer of the mid-18th century, described Prussia as "the most slavish country in Europe," and Carl Philipp Emanuel Bach, the most distinguished of the musicians serving Frederick, did so rather reluctantly. Frederick's religious tolerance, however, was genuine: it was one of the things that helped to mark him in the eyes of contemporaries as a truly enlightened ruler. The abolition of judicial torture, one of his first acts as king, also showed his genuine belief in this aspect of enlightened reform. On an even more fundamental level, the General Education Regulations (General-Landschul-Reglement) of 1763 attempted to create a system of universal primary education throughout the Prussian monarchy. Lack of resources limited its practical effect, but it was the most ambitious effort of the kind theretofore seen anywhere in Europe.

Enlightened reforms

Significance of Frederick's reign. Both by his accomplishments and by his example Frederick deeply influenced the course of German history. In the struggles of the 1740s and '50s he weakened still further the tottering structure of the Holy Roman Empire. The bitter Austro-Prussian rivalry that he began was to be a dominant political force in Germany and central Europe for well over a century. Not until the final Prussian victory over Austria in 1866 was the long contest for leadership in Germany finally resolved. For his share in creating the division of the German world Frederick was later attacked, sometimes bitterly, by a number of historians who saw him as having prevented the emergence of a united Great Germany that included all the major German-speaking areas of Europe. Certainly, he had no sympathy, and indeed no understanding, for the embryonic German nationalism. The efforts of some writers of the 19th and 20th centuries to present him as a forerunner of German national unity are quite misleading. His renewed attack on Maria Theresa of Austria in 1744, for example, frustrated an Austrian invasion of Alsace and its possible return from French to German control, and during the Seven Years' War he offered more than once to cede to France territory in western Germany in the hope of breaking up the coalition that threatened him. Moreover, by his part in the first partition of Poland he helped to create an important common interest with Russia: thenceforth both states had as one of their major objectives the suppression, or at least the strict control, of Polish nationalist aspirations. For generations to come this was to be a factor turning Prussia's attention to eastern Europe and making it less Western in some of its political attitudes than might otherwise have been the case. Yet in many ways Frederick deserved the admiration that later generations, especially in Germany, increasingly felt for him. For all his social and intellectual conservatism he never ceased to feel himself in sympathy with the enlightened intellectual currents and political strivings of the age and with their tolerant and humanitarian aspects. Building on the foundations laid by his father, he consolidated a Prussian ethos of duty, effort, and discipline that, despite some serious negative features, was to become for several generations one of the major political traditions of Europe.

BIBLIOGRAPHY. Frederick's voluminous writings are collected in *Oeuvres de Frédéric le Grand,* 31 vol. in 33, ed. by J.D.E. PREUSS (1846–57), while selections from his correspondence, *Politische Correspondenz Friedrichs des Grossen,* 47 vol. (1879–1939), offer a mine of information on his foreign policy. Writings on his wars were assembled by the Prussian General Staff in *Die Kriege Friedrichs des Grossen,* 19 vol. in 18 (1890–1914). Among the biographies, THOMAS CARLYLE, *History of*

Friedrich II of Prussia, Called Frederick the Great, 6 vol. (1858–65, reissued in 8 vol., 1974), is famous and comprehensive but disappointing in its hero worship. J.D.E. PREUSS, *Friedrich der Grosse: Eine Lebensgeschichte,* 5 vol. (1832–34, reprinted in 9 vol., 1981), is an excellent work of official historiography of its time. REINHOLD KOSER, *Geschichte Friedrichs des Grossen,* 4 vol. (1912–14), available in many later editions, remains a standard large-scale treatment, complete on war and diplomacy. ARNOLD BERNEY, *Friedrich der Grosse: Entwicklungsgeschichte eines Staatsmannes* (1934), takes the story only to 1756; G.P. GOOCH, *Frederick the Great, the Ruler, the Writer, the Man* (1947, reissued 1990), offers essays on different aspects of the subject; D.B. HORN, *Frederick the Great and the Rise of Prussia* (1964), is a short introduction; and PETER PARET (ed.), *Frederick the Great: A Profile* (1972), collects useful extracts from German periodicals and books. A perceptive general discussion is presented in GERHARD RITTER, *Frederick the Great: A Historical Profile* (1968, reprinted 1974; originally published in German, 1954). RUDOLF AUGSTEIN, *Preussens Friedrich und die Deutschen* (1968, reprinted 1986), is extremely hostile but stimulating.

Frederick's early years are described in the old but still useful ERNEST LAVISSE, *The Youth of Frederick the Great* (1892, reissued 1972; originally published in French, 1891). HUBERT C. JOHNSON, *Frederick the Great and His Officials* (1975); and WALTHER HUBATSCH, *Frederick the Great of Prussia: Absolutism and Administration,* trans. from German (1975), study his regime in Prussia. Other special studies include W.O. HENDERSON, *Studies in the Economic Policy of Frederick the Great* (1963); CHRISTOPHER DUFFY, *The Army of Frederick the Great* (1974); HERBERT BUTTERFIELD, *The Reconstruction of an Historical Episode: The History of the Enquiry into the Origins of the Seven Years' War* (1951), a summary of the historical controversy over the attack on Saxony in 1756; HERBERT H. KAPLAN, *The First Partition of Poland* (1962, reissued 1972), analyzing Frederick's share in the dismemberment of the Polish republic; and two works focusing on the conflict over Bavaria: HAROLD TEMPERLEY, *Frederick the Great and Kaiser Joseph: An Episode of War and Diplomacy in the Eighteenth Century,* 2nd ed. (1968); and PAUL P. BERNARD, *Joseph II and Bavaria: Two Eighteenth Century Attempts at German Unification* (1965).

(M.S.An.)

French Literature

Since the Middle Ages, France has enjoyed an exceptional position in European intellectual life. Though its literary culture has no single figure comparable to Dante in Italy or Shakespeare in England, its writers and their language have exercised influence far beyond its borders. In medieval times, because of the political link with Britain, the universality of Latin, and the similarities of the languages derived from Latin, there was a continual process of exchange, in form and content, among the literatures of western Europe. The evolution of the nation-states and the rise in prestige of the vernaculars gradually eroded this relationship, however, and France developed a cultural tradition based on the imitation of Classical models.

The French tradition prized reason, formal perfection, and purity of language and was admired for its thinkers as much as for its writers. By the end of the ancien régime the logic of Descartes, the restraint of Racine, and the wit of Voltaire were seen as the hallmarks of French culture and were emulated throughout the Continent.

The political and intellectual revolutions at the end of the 18th century led to a revaluation. From the 1820s Romanticism openly challenged the Classical ideal and asserted the claims of the imagination against reason and the individual against the social norm. The 12-syllable alexandrine remained the standard line in verse but the form was relaxed, and the domain of poetry was extended successively by Victor Hugo, Charles Baudelaire, and Arthur Rimbaud. All poetic forms were redefined as a result of the Modernist revolutions of the 20th century. In the novel, the dominant literary form from the 19th century, French writers explored the possibilities of the genre, from the novel cycle of Honoré de Balzac to the social realism of Émile Zola. But in the work of writers as different as Stendhal, Gustave Flaubert, and Marcel Proust an analytical approach to questions of style and human motivation seemed both to characterize the French mind and to form a link with the French Classical tradition.

During the first half of the 20th century, Paris was the hub of European intellectual and artistic life, but its position was challenged after World War II, and the international status of the French language declined steadily. Nevertheless, French was the medium of expression for many in Europe, North America, Africa, and Asia, and foreign Francophone (French-speaking) authors contributed significantly to French culture. The *nouveau roman*, or new novel, and such later movements as deconstruction mounted a radical attack on genres, the relationship between writer and reader, and language itself. This, too, was consistent with a cultural tradition notable for its intellectual rigour as well as its appreciation of literary style. (R.C.Bu./Ed.)

This article surveys French literature from the 9th century (to which the earliest surviving fragmentary texts belong) to the present day. See also the section on literature in French in the article BELGIAN LITERATURE and, for the medieval period, the *Early Middle English* section of ENGLISH LITERATURE. Literary works written in French by Canadian authors are discussed in CANADIAN LITERATURE: *Canadian literature in French.*

The article is divided into the following sections:

The Middle Ages

EARLY OLD FRENCH LITERATURE

The origins of the French language. By 50 BC, when the Roman occupation of Gaul under Julius Caesar was complete, the region's population had been speaking Gaulish, a Celtic language, for some 500 years. Gaulish, however, gave way to the conquerors' speech, Vulgar Latin, which was the spoken form of the non-Classical Latin used by the soldiers and settlers throughout the Roman Empire. In different regions, local circumstances determined Vulgar Latin's evolution into the separate tongues that today constitute the family of Romance languages, to which French belongs. This linguistic development was speeded by the empire's collapse under the impact of the 5th-century barbarian invasions. Gaul was overrun by Germanic tribes, in the north principally by the Franks (who gave France its name), and by the Visigoths in the south. But the Latin speech survived: it not only was the language of the majority of the population but also was backed by its associations with the old Roman culture and with the new Christian religion, which used Low Latin, its own form of the Roman tongue. Although it retained relatively few Celtic words, the developing language had its vocabulary greatly enriched by Germanic borrowings, and its phonetic development was influenced by Germanic speech habits. The 9th-century Norse incursions and settlement of Normandy, however, left few traces in the language.

The Romans had introduced written literature, and until the 12th century almost all documents and other texts were in Latin. The first text in recognizable Old French is the Romance version of the Oath of Strasbourg (842), an oath sworn by Louis the German and Charles the Bald against their brother Lothair in the partitioning of the empire of their grandfather Charlemagne. A German version also survives. Only a few other texts, all religious in content, survive from before about 1100.

The Oath of Strasbourg

By about the 9th century a broad division had emerged between the speech of the north of Gaul, which had suffered most from the invasions, and that in the more stable, cultured, and linguistically conservative south. The tongue spoken to the north of an imaginary line running roughly from the Gironde River to the Alps was the *langue d'oïl* (the future French), and to the south was spoken the *langue d'oc* (Provençal or Occitan), terms derived from the respective expressions for "yes."

Vulgar Latin's development had not been uniform throughout the area of the *langue d'oïl;* and, by the time a recognizable Old French had developed, various dialects had evolved, notably Francien (in the Île-de-France, the region around Paris), Picard, Champenois, and Norman. From the latter stemmed Anglo-Norman, the French used alongside English in Britain, especially among the upper classes, from even before the Norman Conquest (1066) until well into the 14th century. Each dialect had its own literature. But for various reasons the status of Francien increased until it achieved dominance in the Middle French period (after 1300); and from it Modern French developed. Old French was a fine literary medium, enlarging its vocabulary from other languages such as Arabic, Provençal, and Low Latin. It had a wide phonetic range and, until the decay of its two-case system inherited from Latin, had syntactic flexibility.

The context and nature of French medieval literature. Whatever Classical literature survived the upheavals of the early Middle Ages was preserved, along with pious Latin works, in monastic libraries. By encouraging scholars and writers, Charlemagne had increased the Latin heritage available to educated vernacular authors of later centuries. He also left his image as a great warrior-emperor to stimulate the legend-making process that generated the Old French epic. There one finds exemplified the feudal ideal, evolved by the Franks to combat social fragmentation and insecurity. The warrior's code of morality, founded on loyalty, bolstered the entire political system. As stability increased under the Capetians, windows opened onto other cultures and elements: that of the Arabs in Spain and, with the Crusades, the East; the advanced Provençal civilization; and the legends of Celtic Britain. The Roman Catholic Church grew in wealth and power, and by the 12th century its schools were flourishing, training generations of clerks in the liberal arts. Society itself became less embattled, and the nobility became more leisured and sophisticated. The muscularity of the epics was tempered by the social graces of *courtoisie:* generosity, modesty, and consideration for others, especially the weak and distressed, and with love regarded no longer as a weakness in a knight but as an objective inspiring and not hindering chivalry.

By the 13th century an additional source of patronage for writers and performers was the bourgeoisie of the developing towns. New genres emerged, and, as literacy increased, prose found favour as a less frivolous medium than verse. There is in much of the literature a rather irreverent spirit, a sometimes cynical realism, yet, at the same time, a countercurrent of deep spirituality. In the 14th and 15th centuries France was ravaged by war, plague, and famine. Alongside a preoccupation in literature with death and damnation appeared a contrasting refinement of expression and sentiment bred of nostalgia for the courtly, chivalric ideal. At the same time a new humanistic learning anticipated the coming Renaissance.

Oral "literature" of the jongleurs

Before 1200 almost all French "literature" had been in verse and had been communicated orally to its public. The jongleurs, professional minstrels, travelled and performed their extensive repertoires, which ranged from epics to the lives of saints (the lengthy romances were not designed for memorization), sometimes using mime and musical accompaniment. Seeking an immediate impact, most poets made their poems strikingly visual in character, more dramatic than reflective, and revealed psychology and motives through action and gesture. Verbal formulas and clichés were used by the better poets as an effective narrative shorthand, especially in the epic. Such oral techniques left their mark throughout the period.

The chansons de geste. More than 80 chansons de geste ("songs of deeds") are known, the earliest and finest being the *Chanson de Roland* (*c.* 1100). Most are anonymous and are composed in lines of 10 or 12 syllables grouped into *laisses* (tirades) based on assonance or, later, rhyme. Their length varies from about 1,500 to more than 18,000 lines. They are exemplary stories of warfare, often pitting Franks against Saracens, that fire the emotions with their insistent rhythms. Under the influence of the genre known as romance, however (see below), the chansons de geste lost some of their early vigour. Their story lines became looser, their adventures more exotic, and their tone often amatory or even humorous. Many were eventually turned into prose.

Cycles formed as new songs were composed featuring heroes, families, or themes already familiar. The *Chanson de Roland* belongs to the cycle known as the "Geste du Roi," or "Geste of the King," the king being Charlemagne, Roland's uncle, in whose service he perished with the rear guard at Roncevaux. Charlemagne is treated less reverently in the parodic *Pèlerinage de Charlemagne* ("Charlemagne's Pilgrimage"), describing his journey to Jerusalem and Constantinople. Dominating the "Geste de Garin de Monglane" is Garin's great-grandson, Guillaume d'Orange, whose historical prototype was count of Toulouse and Charlemagne's cousin. His dogged loyalty to an unworthy monarch (Charlemagne's son Louis) is the subject of a group of poems including the *Chanson de Guillaume* ("Song of William"). The epics in the "Geste de Doon de Mayence" deal with rebellious vassals, among them Raoul de Cambrai, in a gripping story of injustice and strained loyalties. The fanciful 13th-century *Huon de Bordeaux,* which introduces the fairy king Auberon (Shakespeare's Oberon), has been placed here and in the "Geste du Roi." The First Crusade is handled, with legendary embellishment, in a minor cycle.

Controversy surrounds the origins of the genre. It is not known how most of the poems came to contain elements, somewhat garbled, from Carolingian history some 300 years before their composition. Some scholars believe in a continuous process of oral transmission and elaboration culminating in the writing down of the epics as they have survived. Others suppose the historical facts were retrieved much later by poets wishing to celebrate certain heroes, many of whom were associated with pilgrim routes which the jongleurs could then ply with profit. Some evolutionary process seems probable; yet the author of the *Chanson de Roland* (perhaps the Turoldus named in the last line) was undoubtedly a poet of both genius and learning.

The romance. The romance, which came into being in the middle of the 12th century in France, was a creation of formally educated poets. The earliest took their subjects from antiquity: Alexander the Great, Thebes, Aeneas, and Troy were all treated at length, and shorter *contes* ("tales") were derived from Ovid. Other romances, like *Floire et Blancheflor* (adapted in Middle English as *Flores and Blancheflur*), exploited Greco-Byzantine sources; but by about 1150 the Celtic legends of Britain were capturing the public's imagination.

The romance's standard metre is octosyllabic rhyming couplets. It differs from the chanson de geste in concentrating on individual rather than communal exploits and presenting them in a more detached fashion. It offers fuller descriptions, freer dialogue, and more authorial intervention. Christian miracles and fervour are replaced by Eastern or Celtic marvels and the cult of *courtoisie* and *amour courtois* ("courtly love"). There is more interest in psychology, especially in the love situations.

The universally popular legend of Tristan and Iseult had evolved by the mid-12th century, apparently from a fusion of Scottish, Irish, Cornish, and Breton elements. The main French versions (both fragmentary) are by the Anglo-Norman poet Thomas (*c.* 1170) and the Norman Béroul (rather later and possibly composite). The legend was reworked in French prose and widely translated (Thomas' version can be reconstructed from Gottfried von Strassburg's German rendering and another in Old Norse). Chrétien de Troyes's treatment, mentioned in his *Cligès,* has been lost.

Arthurian romance

The deep-rooted British tradition of King Arthur was firmly established on the Continent by Geoffrey of Monmouth's *Historia regum Britanniae* (*c.* 1136; *History of the Kings of Britain*), translated and romanticized by the Jerseyman Wace as the *Roman de Brut* (1155). The Bretons and Anglo-Normans were likely intermediaries in the transmission of further Arthurian material to French writers such as Chrétien de Troyes, the virtual founder of Arthurian romance, who wrote between about 1165 and 1182. His first known romance, *Erec et Enide,* is a serious study of marital and social responsibilities and contains elements of Celtic enchantment. *Cligès,* a partly Greco-Byzantine tale of young love and an adulterous relationship, uses the motif of feigned death later familiar from *Romeo and Juliet. Lancelot,* or *Le Chevalier de la charrette,* relates the infatuated hero's rescue of the abducted queen Guinevere. *Yvain,* or *Le Chevalier au lion,* treats the converse of the situation depicted in *Erec.* Chrétien's ironies and ambiguities invited divergent interpretations, and of no work more than the incomplete *Perceval,* or *Le Conte du Graal,* which may be the conflation of two unfinished poems. The grail, first introduced here, was to become, as the Holy Grail, a remarkably potent symbol. The romance genre was diversely exploited well into the 14th century, but by 1384 Jean Froissart's contribution, *Méliador,* was only a ponderous valediction to romance's golden age. On the genre's periphery were short courtly tales and lais like those of Marie de France, treating Celtic themes and probably composed in England. The unique *Aucassin et Nicolette,* a charmingly comic idyll told in alternating sections of verse (to be sung) and prose (to be recited), pokes sly fun at the conventions of epic and romance alike.

Lyric poetry to the 13th century. The 12th century saw the revolution in sexual attitudes that has come to be known as *amour courtois,* or courtly love. Its first exponents were the Provençal troubadours, poet-musicians of the 12th and 13th centuries, of whom some 400 are known by name. Among them are nobles of both sexes and even clerics. The troubadours no longer considered women to be the disposable assets of men. On the contrary, the enjoyment of a woman's love was a man's aspiration, achievable, if at all, only after the suitor had served a period of amorous vassalage modelled on the subject's service to his lord. This is the main theme of the troubadours' songs, whose origins have been sought in Arabic poetry, the writings of Ovid, Latin liturgical hymns, and other less likely sources. The *canso* (French *chanson*), made of five or six stanzas with a summary envoi, was the favourite vehicle for their love poetry; but they used various other forms, from dawn songs to satirical or debating poems, all usually highly crafted. Guillaume IX, duke of Aquitaine, the first known poet in the Provençal language, mixed obscenity with his courtly sentiments. Among the finest are the graceful Bernard de Ventadour; Jaufré Rudel, who expressed an almost mystical longing for a distant love; and the soldier and poet Bertran de Born.

The *langue d'oïl* had a tradition of dance and spinning songs before the troubadours exerted by the mid-12th century an influence encouraged by, among others, Eleanor of Aquitaine, granddaughter of Guillaume IX and queen of France and later England. The troubadours' verse inspired a number of northern trouvères, including Chrétien de Troyes, two of whose songs are extant, and some nobles such as Thibaut (Theobald I), count of Champagne and king of Navarre.

More interesting is the work of certain bourgeois poets, notably, in the 13th century, a group from Arras and especially Rutebeuf, a Parisian often compared with Villon. Rutebeuf wrote very personal verse on a variety of subjects: his own pitiful circumstances, the quarrel between the University of Paris and the religious orders, the need to support the Crusades, his reverence for the Virgin, and his disgust at clerical corruption. But he did not treat love, nor was his verse set to music.

Satire, the fabliaux, and the Roman de Renart. Medieval literature in both Latin and the vernacular is full of sharp, often bitter criticism of the world's evils: the injustice of rulers, churchmen's avarice and hypocrisy, corruption among lawyers, doctors' quackery, and the wiles and deceits of women. It appears in pious and didactic literature and, as authorial comment, in other genres but more usually in general terms than as particular, corrective satire. Human vice and folly also serve purely comic ends, as in the fabliaux. These fairly short verse tales, most of which are anonymous though some are by leading poets, generate laughter from situations extending from the obscene to the mock-religious, built sometimes around simple wordplay and frequently around elaborate deceptions and counter-deceptions. They are played out in all classes of society but predominantly among the bourgeoisie. Many fabliaux carry mock morals, inviting comparison with the didactic fables. Realistic in tone, they paint instructive pictures of everyday life in medieval France. After the 13th century, when the majority were composed, they yielded in importance to the farces, bequeathing a fund of anecdotes to later writers such as Chaucer and Boccaccio.

Inspired partly by the popular animal fable and more specifically by the Latin poem *Ysengrimus* by Nivard of Ghent (*c.* 1150), Pierre de Saint-Cloud composed about 1175 a poem chronicling the rivalry of Reynard the Fox and the wolf Ysengrin, or Isengrin. This seems to have prompted other writers to relate more of the lively and largely scandalous goings-on in the animal kingdom ruled by Noble the Lion. By the 14th century about 30 branches existed, forming a veritable beast epic. Full of close social observation, they exude the earthy humour of the fabliaux; but, particularly in some of the later branches, this is sharpened into true satire directed against abuses in church and state, with the friars and rapacious nobility as prime targets.

Allegory. Allegory, popular from early times, was employed in Latin literature by such authorities as Augustine, Prudentius, Martianus Capella, and, in the late 12th century, Alain de Lille. It was used widely in religious and moralizing works, as in the long *Pèlerinage de la vie humaine* ("Pilgrimage of Human Life") by Guillaume de Deguileville, Dante's contemporary and a precursor of Bunyan. But the most influential allegorical work in French was the *Roman de la Rose* (*Romance of the Rose*), where courtly love is first celebrated, then undermined. The first 4,058 lines were written about 1230 by Guillaume de Lorris, a sensitive, elegant poet who, through a play of allegorical figures, analyzed the psychology of a young couple's venture into love. The affair is presented as a dream, in which the plucking of a crimson rose by the dreamer/lover would represent his conquest of the lady. Guillaume, however, left the poem unfinished, with the dreamer frustrated and his chief ally imprisoned. Forty years or more later, a poet of very different temperament, Jean de Meun (or de Meung), added more than 17,700 lines to complete it, submerging Guillaume's delicate allegory with debates and disquisitions, laden with medieval and ancient learning, by the characters. Courtly idealism is shunned for a practical, often critical or cynical view of the world. Love, only one of many topics treated in the completed version, is synonymous with procreation; and a misogynistic tone pervades the writing. Embodying these two characteristically medieval but diametrically opposed attitudes to love, the *Roman* was immensely popular until well into the Renaissance (which in some respects is foreshadowed by Jean's humanistic learning) and gave rise to a long-running literary quarrel on the nature and role of women.

The *Roman de la Rose*

Lyric poetry in the 14th century. Allegory and similar conceits abound in much late medieval poetry, as with Guillaume de Machaut, the outstanding musician of his day, who composed for noble patronage a number of narrative *dits amoureux* and a quantity of lyric verse. A talented technician, Machaut did much to popularize and develop the relatively new fixed forms: ballade, rondeau, and virelai. Eustache Deschamps, his great admirer and perhaps also his nephew, struck in his verse a more personal note than many of his contemporaries. A prolific writer, he dealt with public and private affairs, sometimes satirically; but he composed little love poetry, and his work was not set to music. Jean Froissart wrote pleasantly in a variety of lyric forms; but more individual is the varied output of Christine de Pisan, the gifted daughter of an Italian astrologer and physician at the French court. Widowed at 25, she expressed her grief in some of her finest verse. Elsewhere, a prime mover in the controversy provoked by the *Roman de la Rose,* she espouses the feminist cause. Most court verse of this period has an unreal air as if, amid the political and social agonies of the Hundred Years' War, the poets were voicing a yearning for humane and gracious living founded on the ideals of *courtoisie.* Thus Alain Chartier, a political polemicist in both French and Latin, was most admired for *La Belle Dame sans merci,* which tells of the death of a lover rejected by his lady.

Poetry of Christine de Pisan

Villon and his contemporaries. One distinguished victim of the Hundred Years' War was Charles, duc d'Orléans, who was captured at Agincourt at the age of 21 and was held prisoner in England for 25 years. There is an elegiac tone to much of his graceful verse. On his return to France, his court at Blois became a literary centre. For one of his poetry competitions François Villon produced a ballade that, while not outstanding, is remarkable for its individuality and relative lack of contrivance.

Born in Paris in 1431 as François de Montcorbier, Villon adopted the name of his uncle, a priest, who saw to his upbringing. At the University of Paris, where he became Master of Arts in 1452, he acquired some learning but also loose habits that involved him in manslaughter and robbery. His forced departure from Paris was the occasion for his *Lais,* or *Le Petit Testament* (1456; *The Legacy*). This mock legacy in eight-line octosyllabic stanzas is conversational and often facetious in tone, full of allusions to people and events sometimes made cryptic by Villon's taste for antiphrasis. His main work, the *Testament* (or *Le Grand Testament*), was written five or six years later after a spell in the Bishop of Orléans's dungeons. It uses the octets of the *Lais* interspersed with ballades and rondeaux and is similarly packed with personal gossip, often tongue-in-cheek, but leaving a bitter aftertaste. Following more brushes with justice, he disappeared for good. Commonly considered to have been the first modern French poet, he was more traditional than innovative in his themes. Apart from the *Lais* and *Testament,* he produced a variety of lyrics, none provided with music in his day. His verse shows great technical skill and an economy of expression that not only enhances his lively wit but produces moments of intensely focussed vision and, in individual poems, moving statements of human experience.

None of his contemporaries or immediate successors was able to match the vigour of his verse. Often obsessed by metrical ingenuity, extravagant rhymes, and other conceits, they favoured Italian as well as Classical models, thus heralding the Renaissance. It is unfair, however, to judge them by their words alone, since the music was, for most, a vital ingredient of their art.

Prose literature. Prose flourished as a literary medium from about 1200. A few years earlier Robert de Borron had used verse for his *Joseph d'Arimathie* (recounting the Holy Grail's sacred origins) and his *Merlin;* but both were soon turned into prose. Other Arthurian romances adopted it, notably the great Vulgate cycle with its five branches by various hands. These included the immensely popular *Lancelot,* the *Queste del Saint Graal* (whose Cistercian author used Galahad's Grail quest as a vehicle for the mystic pursuit of Christian truth and ecstasy), and *La Mort le Roi Artu,* powerfully describing the collapse of the Arthurian world. The Tristan legend was reworked and extended in prose. To spin out their romances while maintaining their public's interest, authors wove in many characters and adventures, producing complex interlacing patterns, which Sir Thomas Malory simplified when he drew on them for his *Morte Darthur.*

As well as traditional material, new fictions appeared in prose. Early in the 15th century the ironically titled *Les Quinze Joies de mariage* amusingly pursued the antifeminist line. In *Petit Jehan de Saintré* (1456) Antoine de la Sale took a misogynistic view of an ill-starred courtly love relationship, and the work's realism and psychological interest have made it for some the first French novel. The bawdy tales of the *Cent Nouvelles Nouvelles* (*c.* 1465), loosely modelled on Boccaccio, are more in the spirit of the fabliau, though written for the Burgundian court.

Chronicles

Pious and instructional works abound. More interesting are the chronicles, which avoid the romantic extravagances of their verse predecessors. Geoffroi de Villehardouin's *Conquête de Constantinople* ("Conquest of Constantinople") is a sober, if biassed, eyewitness account of the Fourth Crusade (1199–1204). Jean, sire de Joinville, was 84 when, in 1309, he completed his *Histoire de Saint Louis,* a flattering biographical portrait of his intimate friend Louis IX, whom he had accompanied on the Seventh Crusade. Jean Froissart, who travelled extensively in England and Scotland and on the Continent, projected his admiration of chivalry into his four books of chronicles. Covering the years 1325 to 1400, they contain much picturesque detail, largely from personal observation. A far more cynical view of people, politics, and feudal values is found in the *Mémoires* of Philippe de Commynes, with whom modern French historiography may be said to begin.

Religious drama. Classical theatre having disappeared, serious drama was reborn in the Middle Ages within the Roman Catholic Church. There, from early times, dramatic elements were introduced into certain offices, particularly at Easter and Christmas. From this practice liturgical drama sprang. Performances took place inside churches, with the cast of clergy moving from place to place in the sanctuary. At first only Latin was used, though occasionally snatches of vernacular verse were included, as in the early 12th-century *Sponsus* ("The Bridegroom"), which uses the Poitevin dialect. Stories from the Bible and lives of the saints were dramatized; and as the scope of the dramas broadened, more plays were performed outside the church and used only the vernacular. The all-male casts employed multiple setting (*décor simultané*) and moved from one setting, or mansion, to another as the action demanded.

The first extant *mystère,* or mystery play, with entirely French dialogue (but elaborate stage directions in Latin) is the *Jeu d'Adam* ("Play of Adam"). It is known from a copy in an Anglo-Norman manuscript, and it probably originated in England in the mid-12th century. With lively dialogue and the varied metres characteristic of the later *mystères* (all of which were based on Biblical stories), it presents the Creation and Fall, the story of Cain and Abel, and an incomplete procession of prophets. Neither it nor the later *Résurrection du Sauveur* ("Resurrection of the Saviour"; also Anglo-Norman and fragmentary) shows the events preceding the Crucifixion; these first appeared in the early 14th century in the *Passion du Palatinus* ("Passion of Palatinus"). Of relatively modest proportions, this contains diversified dialogue with excellent dramatic potential and probably drew on earlier plays now lost.

The oldest extant *miracle,* or miracle play (a real or fictitious account of the life, miracles, and martyrdom of a saint), is the remarkable *Jeu de Saint Nicolas* ("Play of Saint Nicholas"), by Jean Bodel of Arras, in which exotic crusading and boisterous tavern scenes alternate. Rutebeuf's *Miracle de Théophile* is an early version of the Faust theme, in which the Virgin Mary secures Théophile's salvation. From the 14th century comes the *Miracles de Notre-Dame,* a collection of 40 miracles, partly based on a nondramatic compilation by Gautier de Coincy. These miracles probably were performed by the Paris goldsmiths' guild.

By the 15th century, societies had been formed in various towns for the performance of the increasingly elaborate mystery plays. In Paris the Confrérie de la Passion (Confraternity of the Passion) survived until 1676, though its production of sacred plays was banned in 1548. Notable authors of *mystères* are Eustache Marcadé; Arnoul Gréban, organist and choirmaster at Notre-Dame, and his brother Simon; and Jean Michel. Arnoul Gréban's monumental *Mystère de la Passion* (*c.* 1450, reworked by Michel in 1486) took four days to perform. Other plays took up to eight days. Biblical material was supplemented with legend, theology, and elements of lyricism and slapstick, and spectacular stage effects were employed.

Secular drama. A crucial factor in the emergence of the comic theatre was the oral presentation of much medieval literature. A natural consequence was complete dramatization and collaborative performances by jongleurs and later by guilds or *confréries* formed for the purpose.

Comic plays

The earliest comic plays extant date from the second half of the 13th century. *Le Garçon et l'aveugle* ("The Boy and the Blind Man"), a simple tale of trickster tricked, could have been played by a jongleur and his boy and ranks as the first farce. The Arras poet Adam de la Halle composed two unique pieces: *Le Jeu de la feuillée* ("The Play of the Bower"), a kind of topical revue for his friends, and *Le Jeu de Robin et de Marion,* a dramatized *pastourelle* (a knight's encounter with a shepherdess and her friends) spiced with song and dance. The first serious nonreligious play was *L'Estoire de Griseldis* (1395), the story of a constant wife.

The profane theatre eventually had its own societies of actors, such as the Basoches (associations of lawyers and clerks) and the Enfants sans Souci (probably a special group of Basochiens) in Paris. The societies frequently presented plays in triple bills: first a *sotie,* a slight, sometimes satirical, sketch; next a *moralité* (morality play), a didactic and often allegorical piece; and finally a farce. Some 150 farces have survived from the 15th and 16th centuries. Most are of less than 500 lines and involve a handful of characters acting out plots similar to those of the fabliaux. They use the octosyllabic rhyming couplet and may include songs, commonly in rondeau form. By far the best is the unusually long *La Farce de maistre Pierre Pathelin* (*c.* 1465), a tale of trickery involving a sly lawyer, a dull-witted draper, and a crafty shepherd. Its multifaceted humour is worthy of Molière.

ANGLO-NORMAN LITERATURE

Anglo-Norman is an important and in some respects pioneering branch of medieval French literature. Most genres are represented; and although scholars know of no chanson de geste composed in England, the sole manuscripts of the *Chanson de Roland, Chanson de Guillaume,* and *Pèlerinage de Charlemagne* were copied there. The earliest copy of the *Vie de Saint Alexis* ("Life of Saint Alexis") is likewise Anglo-Norman. The first examples of "scientific" writing in French, from the early 12th century, are the *Cumpoz* or *Comput* (an ecclesiastical calendar), a bestiary, and a lapidary, all allegorical, by a cleric, Philippe de Thaon. He enjoyed patronage at the court of Henry I, as did another cleric and far superior poet, Benedeit, who dedicated his *Vie de Saint Brendan* ("Life of Saint Brendan") to Henry's queen before about 1125. It describes the Irish saint's marvellous voyage in search of Heaven and Hell. Other firsts for Anglo-Norman writers are in the fields of chronicle and drama: Gaimar's verse *Estorie des Engleis* (*c.* 1150), and the *Jeu d'Adam* mentioned above. The first significant prose texts in French seem also to have been produced in England: versions of the Psalter and a translation of the Books of Samuel and Kings.

Anglo-Norman literature is particularly rich in homiletic, devotional, and didactic works, including numerous lives of saints. The trend was encouraged by the church's concern for the instruction of the laity, as expressed by the Fourth Lateran Council (1215). However, a substantial number of chronicles, romances, and even courtly lyrics was produced. Among major writers working in England were Thomas, the author of *Tristan,* and his namesakes Master Thomas, who composed the romance of *Horn,* and Thomas, author of the *Roman de toute Chevalerie,* dealing with Alexander the Great. To Marie de France, author of the lais (see above), are also attributed a collection of fables and an account of St. Patrick's Purgatory, translated from Latin. John Gower, in the late 14th century, was the last notable Englishman to write in any form of French. (D.D.R.O.)

The 16th century

LANGUAGE AND LEARNING IN 16TH-CENTURY EUROPE

The literature of the 16th century does not date precisely from 1500 or 1501. The writers who are identified with Renaissance humanism (a term adapted in France in the 19th century from the German word *Humanismus*) had certain aspirations in common, but they never formed a school. They thought of themselves as the heirs of their predecessors: of Marsilio Ficino, for example, who died in Italy in 1499 but in his lifetime spread Neoplatonic thought widely in Europe, or of Petrarch, who wrote in the 14th century. On the other hand, the period called the Middle Ages in some aspects continued well into the Renaissance, judging by certain titles and subjects of works (even those to which François Rabelais, so "Gothic" in 1532, referred to produce his *Pantagruel*).

Many of the thinkers and writers of the 16th century belong to Europe as a whole more than to a particular nation. This is true of Erasmus, who came originally from Rotterdam but lived in France, England, and Switzerland. The assignment of Jean Lemaire de Belges to a particular country is equally difficult, for he was a Walloon who wrote in French and travelled among various courts. During this period writers made many journeys, either by choice or by necessity. Rabelais, Joachim du Bellay, and Michel de Montaigne all made the trip from France to Italy. Clément Marot died in Turin and Marc-Antoine de Muret, after a long exile, died in Rome. The sometimes intense cultural exchanges (which can be pondered at the crossroads city of Lyon, turned as much toward Italy as toward Paris) explain the influences sometimes submitted to, sometimes acted upon. An example is the rapidity and importance of the diffusion of Martin Luther's writings through France during the decades 1520–40. And the Geneva Reformation made excellent use of Geneva's publishing trade to introduce Luther's ideas in John Calvin's native France. At the time of the Reformation religious literature belonged to all of Europe.

Cultural influences and diffusion

THE ELEVATION OF THE FRENCH LANGUAGE

The communication of ideas was facilitated by the use of Latin, which remained the language of theologians, philosophers, and jurists. Erasmus polemicized in Latin with the Sorbonne or with Luther. Calvin used Latin to write his first version of *Institutio christianae religionis* (1536; definitive Latin version, 1559; *Institutes of Christian Religion*). Petrus Ramus (Pierre de la Ramée) created a sensation when, after earlier writings in Latin, he produced his *Dialectique* (1555) in the vernacular French.

La Pléiade. The group known as La Pléiade was formed in the mid-16th century by seven poets who wished to elevate the French language to the level of Classical languages. Du Bellay's *Deffence et illustration de la langue francoyse* (1549; *The Defence and Illustration of the French Language*) served as La Pléiade's manifesto. In response in part to du Bellay's plea on behalf of French, in the second half of the century a great number of scholarly works were written in the vernacular language. Among them were *Les Dialogues philosophiques,* by Pontus de Tyard, and *Six livres de la République* (1576; *The Six Books of a Commonweale*) by Jean Bodin. Latin did not disappear, however, from the literature of ideas.

The grands rhétoriqueurs. Not all of those mistakenly labelled *grands rhétoriqueurs* ("great rhetoricians") belong chronologically to the 16th century. But Guillaume Crétin, Jean Marot, Jean Bouchet, and above all Lemaire de Belges are all sons of the Renaissance. They left considerable works, sometimes by circumstance (many worked at one time or another as official court historians) and sometimes by desire. As historians, they were "seekers of an origin," to quote Claude-Gilbert Dubois. For example, Lemaire de Belge's *Illustrations de Gaule et singularitez de Troye* (*c.* 1510; "Illustrations of Gaul and Singularities of Troy") tells the legendary origin of the French people, claiming that they are the issue of Francus, the son of

Hector who escaped from Troy. The historians cultivated a sumptuous prose as the expression of their wish for permanence and a desire for glory. The poetry of the *grands rhétoriqueurs* was often sacrificed for circumstance, as seen in Clément Marot, in *Le Voyage de Gênes* (1507) and *Le Voyage de Venise* (1509), and in Lemaire de Belges, in *Le Temple d'honneur et de vertus* (1504) and *La Plainte du désiré* (1509). The poets were fervent partisans of the French language, and the most enthusiastic was Lemaire de Belges, who dared to compare French with Italian in *Concorde des deux langages* (1513). Strictly speaking, it is less their poetic art than their innovations in form that continue to interest scholars.

MAJOR AUTHORS AND INFLUENCES

Poetry. Without the *rhétoriqueurs* the art of Clément Marot could not have been possible. From them, at least at the beginning of his career, he took his inspirations and borrowed the forms to express it, as in the allegorical poem "Le Temple de Cupidon" or in "L'Épître de Maguelonne." But a certain humanist culture; the life at court, which he called his schoolmistress (he became *valet de chambre* to Francis I in 1527); and, above all, the events of his day gave his works a new dimension. As the author of occasional poems Marot preferred the epistle because of its freedom of style and the epigram for its vivacity. With the epistle he reached the summit of the highly subtle art by which he defined himself, a poet of the court but also a Protestant, plagued by the Sorbonne and aspiring to a pure and simple happiness of true religious faith. His allegorical satire on justice, *L'Enfer,* written in 1526 after his brief imprisonment on charges of violating Lenten regulations, and his religious beliefs explain his voluntary exile after the Affaire des Placards in 1534. His return to Paris in 1537 made him no more prudent, as he continued his translations of the Psalms (published in the late 1530s and early 1540s), which were a brilliant literary achievement but were suspect in the eyes of the religious authorities. Marot's translation, continued by the Protestant leader Theodore Beza (Théodore de Bèze), became the Huguenot psalter.

Ronsard

Marot had been dead for six years when Pierre de Ronsard made a remarkable entry into the French literary scene with the publication of the first four books of his *Odes* in 1550. In the *Odes* he sings alternately of the great figures of his day and of the Vendôme countryside. Ronsard indicated in his preface, however, that he started work on his poems at the same time Marot was translating the psalter. The new "school," which at first was merely a group of friends calling itself La Brigade (the name La Pléiade coming into use later), deliberately diverged from the paths taken by their predecessors. The group broke less, on the other hand, with the Lyon school, whose members included Maurice Scève (*Délie,* 1544), Pernette du Guillet (*Rymes,* 1545), and Louise Labé (*Oeuvres,* 1555). Scève, above all, left a very important literary work in *Délie,* in which he mixed Platonism and Petrarchianism in a purposefully difficult language. The poets of La Pléiade were not content to sing of love, the favourite theme of the Lyon poets. True humanists were familiar with antiquity, having read and translated Greek as well as Latin, and antiquity opened to them a vast horizon. They did publish numerous collections of love poems: du Bellay, *L'Olive* (1549); Ronsard, *Les Amours* (1552), *La Continuation des amours* and *La Nouvelle Continuation des amours* (1555–56), and *Les Sonnets pour Hélène* (1578); and Jean-Antoine de Baïf, *Les Amours de Méline* (1552) and *Les Amours de Francine* (1555). But love for them was not only individual sentiment. Love presided over the life of the entire cosmos, as sung by Jacques Peletier in *L'Amour des amours* (1555) and by all the poets influenced by the Neoplatonic view of life. The most celebrated collection by du Bellay, *Les Regrets* (1558), thus draws to a close on a loving vision of beauty. Others engaged in philosophical poetry, of which Ronsard's *Les Hymnes* (1555–56) is the best example.

The spectacle of nature enchanted the poets who, like Peletier, were also learned men. A "scientific" poetry appeared and expressed itself in many ways, from the hymn reinvented by Ronsard to the highly polished verse form chosen by Rémy Belleau in his *Amours et nouveaux échanges des pierres précieuses* (1576). This poetry found its justification in the idea that the poet, because of his inspiration, acceded to knowledge of the world, which conferred upon him a dignity at least equal to that of a priest.

When the civil wars broke out in 1562, the members of La Pléiade sided with the monarchy and the church, which Ronsard eloquently defended in *Discours* (1562–63). Their commitment was guided by a conviction that the French monarchy was the best and that the Roman Catholic Church, subject perhaps to internal reforms, was the true heir of the church established in the first centuries after Christ. If their beliefs at times led them to write cruel verse, notably at the time of the Massacre of Saint Bartholomew's Day (beginning August 23, 1572, when Catholics rampaged and murdered some 3,000 Huguenots in Paris alone), all defended their sincere convictions and, at the same time, certain ideas of literary expression.

Influence of Calvin

The Reformation, in fact, was not limited to dogmatic, liturgical, or ecclesiastical changes. It brought cultural change as well. Ronsard became aware of it when Protestant ministers launched against him a series of scurrilous charges, to which he replied with the *Responce . . . aux injures et calomnies* (1563). Calvin and his disciples allowed neither humanism's global admiration for antiquity nor the poets' taste for mythology. Calvin assigned to the poet an imperative and even exclusive duty: to praise God and serve his Word. This is why Calvin had encouraged Marot's translation of the Psalms and why, after Marot, many Protestant poets, such as Theodore Beza, Jean de Sponde, and Théodore-Agrippa d'Aubigné, paraphrased or pondered them.

Humanist tragedy was born in 1553 with *Cléopâtre captive* by Étienne Jodelle. But before him, in a totally different spirit, Beza had written (and produced in Lausanne) *Abraham sacrifiant* (1550), which retained many qualities of the medieval *mystère* and, using a celebrated biblical episode, illustrated above all the combat of faith. The formula chosen by Beza was taken up again by Loys Desmasures, also a Protestant, in his *tragédies saintes* (1563): *David combattant, David triomphant,* and *David fugitif.* But this formula had no influence on the development of regular tragedy that blossomed in the works of Robert Garnier, published from 1568 to 1580.

Théodore-Agrippa d'Aubigné represented the perfect synthesis of humanism and Calvinism. He studied to perfection the three traditional languages, Latin, Greek, and Hebrew; and he was familiar with modern languages, especially Italian. He was so faithful to his church that he was given the sobriquet "Le Bouc du Désert" ("Goat of the Desert," implying "scapegoat") at a time (during the reign of Henri IV) when religious compromise and the rallying of numerous Protestants rendered his intransigence almost inappropriate. Earlier he had written of love, and he had written verse modelled on Petrarch in *Printemps* (composed between 1568 and 1580). These poems would not be distinguished from so many other collections were it not for the melancholy inspiration and force of their rhetoric. His rhetoric is found, without the melancholy, in his master poem *Les Tragiques,* composed for the most part at the end of the century but not published until 1616. The seven cantos of the poem invited the reader to a theological understanding of history. The first cantos ("Misères," "Princes," and "Chambre dorée") successively depicted a sombre tableau of France exhausted by the civil wars, the debauchery of Henri III's court, and the corrupt and cruel institutions of justice. In the fourth and fifth cantos, "Feux" and "Fers," Aubigné began a long list of the Protestant martyrs and of the fight of their coreligionists in defense of their faith. Aubigné's theological vision is even more accentuated in the last two cantos, "Vengeances" and "Jugement." In "Vengeances" the poet explains the judgments made by God against tyrants and then turns his attention to those who God will lift up at the end of time. "Jugement" is the apotheosis of the poem, the cry of joy of those murdered by history. It is the summit of a work that draws to a close—rare in Protestant works—with a mystical vision.

The term baroque has been used for a style that gives precedence to forms. It would be more just to underline the importance of a rhetoric of persuasion, with all its means: apostrophe, antithesis, descriptions. The pen that wrote *Les Tragiques* is not the same pen that wrote, albeit at the same time, the *Histoire universelle,* in which objectivity regained its hold. With Aubigné flickered the last light of 16th-century poetry. The *Histoire universelle,* his masterpiece, was published in the 17th century but escaped notice. Aubigné was foreign to Classical taste, which Malherbe inaugurated early in the 17th century.

Prose. The production of poetry in the 16th century did not outdo the other genres in quantity. Readers turned above all to works in prose, many of which do not belong to literature as strictly defined: accounts of voyages, lives of saints, and collections of diverse *leçons* or *lectures* (readings). Prose was slow in freeing itself from the heavy yoke thrown over it by the medieval humanists in their concern for the use of language. With Lemaire de Belges prose became eloquent, and with François Rabelais it became a prodigious domain of experimentation.

Rabelais

The work of Rabelais, who Voltaire called a "drunken philosopher," escapes all labels. Humanism rightfully claims *Pantagruel* (1532) and *Gargantua* (1534 or 1535), with their celebrated pages on education, war and peace, and the true religion. They resound with the voice of the man who called Erasmus his spiritual father and who befriended numerous Protestants. But it is important to remember that Rabelais in the first two books of his great work intended to laugh and to give laughter and that nothing, not even the most beautiful humanist philosophy, escaped the corrosive power of laughter.

The humanist reading of the work is embarrassed when it comes to interpreting the last three books, published long after the first two: *Le Tiers Livre* in 1546, *Le Quart Livre* in 1552, and the *Cinquième Livre* in 1564 (and of questionable authenticity). Certain episodes, such as Panurge's posing of the question of marriage, or the voyage of Pantagruel and his companions, often derail the critic who interprets on too high a level.

It is possible, as Verdun-Louis Saulnier thought, that Rabelais tried to deliver an essentially religious message for the world to decode, that of the Protestant humanists who, at the moment of persecution, abandoned the fight for doctrine and took refuge in silence and prayer. It is also quite possible that Pantagruel's voyage is the philosophical symbol for the search for truth against the dogmatism of all orders: dogmas are represented by the islands encountered on the voyage. Perhaps the work of Rabelais makes evident the crisis of pre-Cartesian philosophy, which oscillates between the for and the against, the *doxa* and the paradox, without arriving at the desired synthesis. But if the philosophy is blocked, the language is liberated, alternately solemn and clownish, clear and obscure.

Rabelais was acquainted with the milieu of learned men and their jokes and delighted in multiplying their ambiguity, their implications, and their wordplay. He produced a unique work that no one could imitate. Contemporaries did not err in calling Rabelais a poet.

Montaigne

Michel de Montaigne ranked Rabelais among the "simply pleasant" authors. Indeed, the purpose of Montaigne's *Essais* (*Essays*) had little in common with any of Rabelais's intentions. This is not to say that Montaigne disdained laughter and, above all, humour. But if he retired to the tower of his château while still young, it was to write something other than buffoonery. He too produced adventure novels, but he also invented a new genre, that of the *Essais.* At first he devoted his time to putting down on paper all the thoughts inspired by his numerous readings. But soon he took an interest in himself and became the material of his own book. The first two volumes of the *Essais* were published in 1580. A third was added in 1588, along with an enlarged edition of the first two. When he died in 1592 he left his own copy of the *Essais,* with numerous revisions written in his own hand. This revised text was published in 1595. Montaigne had stated that he could have continued to write indefinitely. The subject given him—his own self—was fathomless, as each day revealed anew, and introspection was endless.

Montaigne rejected any form of dogmatism (though leaving aside the question of the truths of religion). He was constantly on the watch for whatever he could learn about man, from conversation or from books, and especially from books on history and accounts of travels. He declared himself to be a Pyrrhonist, not a Skeptic, because the error of the Skeptic is to say that he knows nothing, while the Pyrrhonist refuses to make even that affirmation.

The essay as Montaigne conceived it adapts itself perfectly to that philosophy. Free in form, the essay permits the author to pass from one idea to another at will, according to his whim, and to the despair of those who seek a plan in his chapters. If there is no plan, it is because Montaigne invented a spontaneous style of writing that invites spontaneity in cognizant readers as well.

In contrasting ways Montaigne and Aubigné both enlightened the end of the 16th century. Aubigné was a party man, a herald of religion, and he deliberately put his poetry at the service of these. Montaigne was a man without party who invented a literature for the *honnête homme* ("gentleman"; the ideal of refinement), which may explain his success in the 17th century. But the qualities represented by Aubigné do not disappear in the 17th century, "the century of saints," which saw the birth of many great religious works. The richness of the 16th century is found also in its legacy. (D.Mé.)

The 17th century

LITERATURE AND SOCIETY

Refinement of the language. At the beginning of the 17th century the full flowering of the Classical manner was still remote, but various signs of a tendency toward order, stability, and refinement can be seen. A widespread desire for cultural self-improvement is reflected in the numerous manuals of politesse, or formal politeness, which appeared through the first half of the century; while at the celebrated salon of Mme de Rambouillet men of letters, mostly of bourgeois origin, and the nobility and leaders of fashionable society mixed in an easy relationship to enjoy the pleasures of the mind. Such gatherings did much to refine the literary language and also helped to prepare a cultured public for the serious analysis of moral and psychological problems.

The earliest imaginative literature to reflect the new taste was written in imitation of the pastoral literature of Italy and Spain; the masterpiece of the genre was *L'Astrée* (1607–27) by Honoré d'Urfé. Manners are stylized, settings are conventional, and the plot is highly contrived; but the psychology of the characters is handled with insight.

Influence of Malherbe

Refinement of the language of poetry was the self-imposed task of François de Malherbe: resolutely opposed to the exalted conception held by La Pléiade of the poet as inspired favourite of the Muses, he owes his place in literary history not to his undistinguished creative writing but to a critical doctrine imposed on fellow poets by word of mouth and personal example. Malherbe called for a simple, harmonious metre and a sober, almost prosaic vocabulary, pruned of imaginative poetic fancy. His influence helped to make French lyric verse, for nearly two centuries, elegant and refined but lacking imaginative inspiration. Malherbe's alexandrine, however—clear, measured, and energetic—was a metre marvellously suited to be a vehicle for Pierre Corneille's dramatic verse.

Not all poets of the 1620s accepted Malherbe's lead. The most distinguished of the independents was Théophile de Viau (referred to as Théophile), who not only opposed Malherbe in style and technique but also expressed the free thought inherited from Renaissance Italy. Théophile's verse, with its engaging flavour of spontaneity and sincerity, shows a sensual delight in the natural world. His whole way of life was a provocation to the *bien-pensants* ("right-minded"): he was the leader of a freethinking bohemia of young noblemen and men of letters, practicing and preaching social and intellectual unorthodoxy. His persecution, imprisonment, and early death ended all this, however: libertinage went underground, and repressive orthodoxy was entrenched for a century or more. The poetry of Théophile and other independents exemplifies that manner to which modern criticism has given the name baroque. The baroque poet has been said to possess "a vision of life distorted through imagination and sensibility." Whereas in the case of Malherbe and his school sensibility is constantly controlled by common sense, baroque writers embellish their descriptions of nature by subjective flights of fancy that may even assume an absurd or surrealist flavour, resulting in an intensely personal picture of the world, far removed from Malherbe's cliché-like generalizations.

Development of drama. Unlike the humanist playwrights of previous generations, Alexandre Hardy was nothing if not a man of the theatre. *Poète à gages* (staff poet) to the Comédiens du Roi company at the Hôtel de Bourgogne in Paris, he wrote hundreds of plays, of which 34 were published (1623–28). In addition to writing tragedies, he developed the tragicomedy and pastoral, which became the most popular genres between 1600 and 1630. While Hardy's plays possess the vigour and colour of Elizabethan and Jacobean drama, his style is unattractive; but, in the theatre as elsewhere, the pastoral was a refining influence, providing a vehicle for the subtle analysis of feeling. Although the finest play of the 1620s is a tragedy, Théophile's *Pyrame et Thisbé* (1623), which shares the fresh, lyrical charm of the pastorals, tragicomedy is without a doubt the baroque form at its best. Here, the favourite theme of false appearances, the episodic structure, and devices such as the play within the play reflect the essentials of baroque art. During the 1630s a crucial struggle took place between this irregular type of drama and a simpler and more disciplined alternative. Theoretical discussion focussed on the conventional rules (the unities of time, place, and action, mistakenly ascribed to Aristotle), but the *bienséances* (conventions regarding subject matter and style) were no less important in determining the form and idiom the mature Classical theatre was to adopt.

Comedy gained a fresh impetus about 1630; and the new style, defined by Corneille as "une peinture de la conversation des honnêtes gens" ("a painting of the conversation of the gentry"), simply transposes the pastoral into an urban setting. At the same time, ambitious young playwrights competing for public favour and the support of the two Paris theatre companies, the Hôtel de Bourgogne and the Marais, did not neglect other types of drama; and Corneille, together with Jean Mairet, Tristan (François L'Hermite), and Jean de Rotrou, inaugurated "regular" tragedy. But it was some time before Corneille, any more than his rivals, turned exclusively to tragedy. The eclecticism of these years is illustrated by *L'Illusion comique* (1635), a brilliant exploitation of the interplay between reality and illusion that characterizes baroque art.

Corneille's *Le Cid*

The two trends come together in Corneille's theatre in *Le Cid* (1637), which, though often called the first Classical tragedy, was created as a tragicomedy. The emotional range Corneille achieves with his verse in *Le Cid* is something previously unmatched. Contemporary audiences at once recognized the play as a masterpiece, but it was subjected to an unprecedented critical attack. The *querelle du Cid* (quarrel of *Le Cid*) caused such a stir that it led to the intervention of the Cardinal de Richelieu, who referred the play to the judgment of the newly founded Académie Française.

The formation of the Académie, an early move to place cultural activity under the patronage of the state, dates from 1634; examination of *Le Cid* on Richelieu's orders was an exception to its normal functions, which concerned the standardization of the French language. This effort bore fruit in the Académie's own *Dictionnaire* of 1694, though by then rival works had appeared in the dictionaries of Richelet (1680) and Furetière (1690).

A similar desire for systematic analysis inspired Claude Favre, sieur de Vaugelas, also an Academician, whose *Remarques sur la langue françoise* (1647) record polite usage of the time. In the field of literary theory the same rational approach produced the *Poétique* of La Mesnardière (1639) and the Abbé d'Aubignac's *Pratique du théâtre* (1657), both treatises, which strongly influenced the establishment of Classical doctrine being instigated by Richelieu's personal patronage.

Meanwhile another protégé, Jean Chapelain, began in the 1630s to exert an influence similar to that of Malherbe a generation earlier. Although he produced no important doctrinal work and made his mark in the salons and in occasional writings, Chapelain was nevertheless a major architect of Classicism in France. More liberal than Malherbe, he made allowance for that intangible element ("le je ne sais quoi") that rules cannot produce. The *Sentiments de l'Académie* (1638), compiled by Chapelain as a judgment on *Le Cid*, reflect prudent compromise, but one can sense beneath the pedantry of certain comments a genuine feeling for the harmony and regularity that Classical tragedy was to achieve.

The effect of the *querelle du Cid* on Corneille's evolution is unmistakable: all his experimentation was henceforth to be carried out within the stricter Classical formula. A remarkable spell of creative activity produced in quick succession *Horace* (1640), *Cinna* (1640), and *Polyeucte* (1643), which, with *Le Cid*, represent the playwright's highest achievement: a triumphant justification of the formula that Mairet and others had helped to develop but which Corneille himself perfected. The essence of Classical tragedy is a single action, seized at crisis point. Despite the prominence always given to the unities of time and place, it is unity of action that gives Classical tragedy its essential character. The other unities merely help to make unity of action effective.

Tragicomedy lingered on as a popular alternative. Rotrou's *Saint-Genest* (1647), for example, provides an interesting contrast with *Polyeucte*, treating in the baroque manner similar themes of divine grace and conversion. But by the 1640s writers and their public had become more responsive to various standardizing influences. René Descartes's *Discours de la méthode* (1637; *Discourse on Method*), with its opening sentence, "Le bon sens est la chose du monde la mieux partagée . . ." ("Good sense is of all things in the world the most equally distributed . . ."), clearly assumes that the mental processes of all men, if properly conducted, will lead to identical conclusions. A similar assumption is implicit, as regards the psychology of the passions, in Descartes's *Traité des passions de l'âme* (1649; *Treatise on Passions*). In the field of creative writing, poets sometimes come to distrust their individual sensibilities and prefer to mold their imaginations to the common denominator of a social group. In lyric poetry, the linguistic tendencies crystallized by the reforms of Malherbe and Vaugelas combine with the preoccupations of the salons to produce writing that is seldom more than mannered wordplay.

Generally speaking, such tendencies toward a literature expressing the cultural values of a homogeneous society affect form more than content. For the self-centred aristocratic idealism that inspired the Fronde (a series of civil wars between 1648 and 1653) also finds expression in the literature of the period, and nowhere more clearly than in Corneille's tragedies. His self-reliant heroes, meeting every challenge and overcoming every obstacle, are motivated by the self-conscious moral code that animated the Cardinal de Retz, Mme de Longueville, and other leaders of the heroic but futile resistance to Cardinal Mazarin. In neither case is devotion to a cause free from self-glorification; in both, the approbation of others is as necessary as the desire to leave an example for posterity. Such optimistic, heroic attitudes may seem incompatible with a tragic view of the world; indeed, Corneille provides the key to his originality in substituting for the traditional Aristotelian emotions of pity and fear a new goal of admiration. Corneille asks that his audience admire something larger than life, and the best of his plays are still capable of arousing this response.

The heroic ideal. The same appetite for heroic subject matter is reflected in the mid-century novels. These resemble *L'Astrée* in that they are long-winded, multivolume adventure stories with highly complicated plots, but they have moved from the world of the pastoral to that of ancient history. The two best known examples, *Artamène ou le grand Cyrus* (1649–53) and *Clélie* (1654–60), both by Mlle de Scudéry, are set in Persia and Rome, respectively. Such novels reflect the society of the time. They also show how the readers and playgoers of the Classical age were formed: the minute analysis of the passions, when divorced from the superficial romanesque, looks forward to the psychological subtlety of Racine.

Other writers of the period make a more individual use of the novel form. Cyrano de Bergerac returned to the Renaissance tradition of fictional travel as a vehicle for social and political satire and may be seen as an early exponent of science fiction. So provocative were the ideas expressed in his *Histoire comique des états et empires de la lune* (1656) and *Histoire comique des états et empires du soleil* (1661; *A Voyage to the Moon: With Some Account of the Solar World*) that neither work was published until after 1655, the year of his death. Paul Scarron was more down-to-earth in purpose and manner: in *Le Roman comique* (1651–57) he set out to parody the heroic novels. His novel follows the fortunes of an itinerant troupe of actors and has some value as documentary, but in his portrayal of provincial society Scarron yields to his penchant for good-humoured burlesque, also illustrated in his mock-epic *Virgile travesti* (1648–52).

The *Maximes* of La Rochefoucauld

The honnête homme. Partly because of the influence of the salons and partly as a result of disillusionment at the failure of the Fronde, the heroic ideal was gradually replaced in the 1650s by the concept of *honnêteté*. The word literally means "honesty" but figuratively refers to a sincere refinement of tastes and manners. Unlike the aspirant after *gloire* ("glory"), the *honnête homme* ("gentleman") cultivated the social graces and valued the pleasures of social intercourse. A cultured amateur, modest and self-effacing, he took as his model the Renaissance *uomo universale* ("universal man"), the complete all-rounder. François de La Rochefoucauld provides an interesting illustration of the transition between the two ages. An aristocrat who played a leading part in the Fronde, he was motivated in his early life by ambition and family pride but retired to begin a new career as man of letters. The *Maximes* (1665), his principal achievement, is a collection of 500 epigrammatic reflections on human behaviour, expressed in the most universal terms: the general tone is cynical, self-interest being seen as the source of all actions. If a more positive message is to be seen, it is the recognition of *honnêteté* as a code of behaviour that holds society together. However, even this is touched with cynicism. La Rochefoucauld's view of *honnêteté* is a pragmatic one, falling as far short of the ideal defined by Antoine Gombaud, chevalier de Méré, in his *Discours de la vraie honnêteté* (1701; "Discourse on True Honnêteté") and other essays, where it is presented as a true *art de vivre,* as it does of the example set by Charles de Saint-Denis, sieur de Saint-Évremond, who, in the opinion of contemporaries, most nearly lived up to such an ideal. Few *honnêtes gens* had the culture, the taste, and the temperament to practice the art of living in such an exemplary way, but the ideal of tolerant, cultured Epicureanism for a while set the tone of social life in Paris.

This period also saw the fullest development of the feminine cult of *préciosité,* a style of thought and expression exhibiting delicacy of taste and sentiment. Inasmuch as *honnêteté* stands for moderation in all things, and *préciosité,* in the extreme, for affectation and extravagance, the two phenomena may seem to be opposites. The sentiments and manners satirized by Molière in *Les précieuses ridicules* (1659) do not represent the whole picture, however, and, although the natural desire of these early feminists to assert themselves meant that their ideas were often taken to extremes, *précieuses* like Mlle de Scudéry were responsible for introducing a new subtlety into the language, establishing new standards of delicacy in matters of taste, and propagating advanced ideas about the equality of the sexes in marriage. Their aims thus ran parallel to those of the *honnêtes gens*, and the ideal of the educated, emancipated woman was the female counterpart of the masculine ideal defined above.

The fullest representation of the *honnête homme* in imaginative literature is to be found in the theatre of Jean-Baptiste Poquelin, known as Molière. A bourgeois by birth, a courtier, and an *honnête homme,* Molière was also an actor-manager and an entertainer. He toured the provinces with his theatre troupe from about 1645 until 1658, when they returned to Paris. Molière soon succeeded in imposing on audiences a completely new type of comedy. While his early plays may be divided conventionally into *comédies littéraires* and popular farces, from *L'École des femmes* (1662) onward these two strains are fused, creating a formula that combined the Classical structure, the linguistic refinement, and the portrayal of manners expected of comedy with the caricatural characterization proper to farce. Even in stylized verse plays such as *L'École des femmes, Le Misanthrope* (1666), *Le Tartuffe* (1664), or *Les Femmes savantes* (1672), the comedy of manners merely provides a framework for the comic portrait of a central character, in which exaggeration and fantasy play a considerable part. However topical the subject and however prominent the satirical content of Molière's plays, his characters always possess a common denominator of universal humanity. Most plays contain, alongside the comic character, one or more examples of the *honnête homme;* and the social norm against which his comic characters offend is that of a tolerant, humane *honnêteté.* In *Le Tartuffe,* and in *Dom Juan* (1665), topical references and satirical implications were so provocative, because both plays dealt with the delicate subject of religious belief, that there were strong reactions from churchmen. However, from the start of his Paris career Molière could count on the active support of the king, Louis XIV. A number of his plays were written for performance at Versailles or other courts; and Molière also wrote several *comédies-ballets* and collaborated in other divertissements that brought together the arts of poetry, music, and dance.

The biggest box-office success of the century, judged by length of first run, was the *Timocrate* (1656) of Pierre Corneille's younger brother Thomas, a prolific playwright adept at gauging the public taste. *Timocrate* was exactly contemporary with the *précieux* novels of Mlle de Scudéry, and, like Philippe Quinault in his *tragédies galantes,* the author reproduced the disguises and amorous intrigues so much admired by habitués of the salons. However, the 1660s were to see the rivalry between two acknowledged masters of serious drama. Pierre Corneille, returning to the theatre in 1659 after a hiatus, wrote several more plays; but though *Sertorius* (1662) or his last play, *Suréna* (1674), bear comparison with earlier masterpieces, the heroic idealism had lost conviction. While Corneille retained his partisans among older playgoers, it was Jean Racine who appealed to newer audiences, less idealistic in their attitude to psychology: the new realism was in tune with La Rochefoucauld's *Maximes* rather than with Descartes's *Traité des passions.*

Racine's fatalism. Whether Racine's Jansenist upbringing determined his view of a human nature controlled by perverse and willful passions, or whether his knowledge of Greek tragedy explains the fatalism of his own plays, the imaginary world inhabited by his heroes could not be more different from that of Corneille's. Tragedy for Racine is an inexorable series of events leading to a foreseeable catastrophe. Plot is of the simplest; the play opens with the action at crisis point; and once the first step is taken, tension mounts between incompatible protagonists until one or more is destroyed. Racine's career began in 1664 with *La Thébaïde,* a grim account of the mutual hatred of Oedipus' sons; this was followed by *Alexandre le grand* (1665), his only attempt at the manner of Quinault. The masterpieces date from the highly successful *Andromaque* (1667), another subject from Greek legend, after which, for *Britannicus* (1669) and *Bérénice* (1670), Racine turned to topics from Roman history. *Bajazet* (1672) is based on modern Turkish history; *Mithridate* (1673) has as its hero the famous enemy of Rome; and finally there followed two plays with Greek mythological subjects: *Iphigénie* (1674) and *Phèdre* (1677).

Writings of Boileau

Nondramatic verse. Nondramatic verse still enjoyed a special prestige, as shown in *L'Art poétique* (1674) of Nicolas Boileau-Despréaux, where the genres most highly esteemed are the epic (of which no distinguished example was written during the century), the ode (a medium for official commemorative verse), and the satire. Boileau himself, in his satires (from 1660) and epistles (from 1674), as well as in *L'Art poétique,* established himself as the foremost critic of his day; but despite a flair for judging contemporaries, his criteria were limited by current aesthetic doctrines. In *Le Lutrin* (1674–83), a model for Alexander Pope's *The Rape of the Lock*, he produced a masterpiece of comic writing in the Classical manner. Jean de La Fontaine's *Fables* (1668; 1678–79; 1692–94) succeed in transcending the limitations of the genre; and although readers formerly concentrated excessively on the moral teaching they offer, it is possible to appreciate beneath their apparent naïveté the mature skills of a highly imaginative writer, who displays great originality in adapting to his needs the linguistic and metrical resources of the Classical age.

THE CLASSICAL MANNER

Though the novel was still considered to be a secondary genre, it produced one masterpiece that embodied the Classical manner to perfection. In *La Princesse de Clèves* (1678) by Mme de La Fayette, the narrative forsakes the fanciful settings of its pastoral and heroic predecessors and explores the relationship between the individual and society in a sober, realistic context. The language achieves its effects by understatement and subtle nuance rather than by rhetorical flourish; and it is appropriate that a woman writer should have created this lasting tribute to the feminine influence, which, in the salons, helped to form such an expressive medium for psychological analysis. The other great woman writer of her age, Mme de Sévigné, was (like La Fontaine) too idiosyncratic to be truly representative. Her intimate, informal correspondence—totally unlike that of Jean-Louis Guez de Balzac half a century earlier—was nevertheless composed with a careful eye to literary effect. Mme de Sévigné not only was an admirable example of the cultured reader for whom the *grands classiques* wrote but was herself one of the most skillful prose writers of her day.

The *Pensées* of Pascal

The most distinguished prose writer of the age, however, was a man who, if he does reflect the society he lived in, does so in a highly critical light. The *Pensées* of Blaise Pascal present an uncompromising reminder of the spiritual values of the Christian faith. The work remains incomplete, so that in place of the dialectical cogency of *Les Provinciales* (1656–57), his masterly satire of Jesuit casuistry, it possesses an enigmatic, incoherent quality in spite of the aphoristic brilliance of many fragments. The central theme is clear enough: Pascal's view of human nature has much in common with that of La Rochefoucauld or Mme de La Fayette, but in his case the misery of godless man is contrasted with the potential greatness man can attain through divine grace. Pascal is the first master of a really modern prose style. Whereas Descartes's prose is full of awkward Latinisms, Pascal uses a short sentence and is sparing with subordinate clauses. The clarity and precision he achieves are equally appropriate to the penetrating analysis of human nature in the *Pensées* and to the irony and comic force of the *Provinciales*.

Religious authors. A new intellectual climate can be recognized from 1680 onward. An increased spiritual awareness resulting from Jansenist teaching, the preaching of Jacques-Bénigne Bossuet and others, and the influence of Mme de Maintenon at court, marked French cultural life with a new moral earnestness and devotion. The position of Bossuet is an ambivalent one. In spite of his outspoken criticism of king and court, his view of kingship and of the relationship between church and state made him one of the principal pillars of the regime of the Sun King, carrying Richelieu's policies to their logical conclusion. His ultraorthodox views are expressed in writings such as the *Discours sur l'histoire universelle* (1681); but he also exerted a considerable moral influence in his sermons and funeral orations, which took the art of pulpit oratory to a new high level. François de La Mothe-Fénelon was a much less orthodox churchman, and the influence he wielded was of a more liberal nature. Like Bossuet, he was a tutor in the royal household, and he was also author of a classical novel, *Les Aventures de Télémaque* (1699).

Just as Fénelon chose a Classical model—his novel purports to be the continuation of Book Four of the *Odyssey*—so Jean de La Bruyère chose to write his *Caractères de Théophraste traduits du grec, avec les caractères ou les moeurs de ce siècle* (1688) in the style of the Greek moralist Theophrastus. However, his work, appended to his translation of Theophrastus, was from the beginning more specific in its reference to his own times; and successive editions, up to 1694, made of it a powerful indictment of the vanity and pretensions of a status-conscious society, and even of the extravagance and warmongering of the King himself. At best, La Bruyère writes as an ironic commentator on the social comedy around him.

An equally satirical picture of the age is left by a number of Molière's successors writing for the comic theatre (which, from the founding of the Théâtre Français in 1680, was organized on a monopoly basis). Comedy, at the hands of such writers as Jean-François Regnard, Florent Carton Dancourt, and Alain-René Lesage, continued to be lively and inventive; but the writing of tragedy, by contrast, already had become a much more derivative exercise. Exception must be made for Racine's last two plays, *Esther* (1689) and *Athalie* (1691), written not for the professional theatre but for the girls' school at Saint-Cyr. The latter in particular is as powerful as any of the secular plays.

The Ancients and the Moderns. Finally, the end of Louis XIV's reign witnessed the critical debate known as the *querelle des anciens et des modernes* (Quarrel of the Ancients and the Moderns), a long-standing controversy that came to a head in the Académie and in various published works. Whereas Boileau and others saw imitation of the literature of antiquity as the only possible guarantee of excellence, moderns such as Charles Perrault in his *Parallèle des anciens et des modernes* (1688–97) and Bernard Le Bovier, sieur de Fontenelle, in his *Digression sur les anciens et les modernes* (1688) claimed that the best contemporary works were inevitably superior, because of the greater maturity of the human mind. It was a sterile and inconclusive debate, but the underlying issue was most important, for the moderns indirectly, if not explicitly, anticipated those 18th-century thinkers whose rejection of a single universal aesthetic in favour of a relativist approach was to hasten the end of the Neoclassical age. (W.D.H.)

The 18th century: from Louis XIV to the Revolution

THE ENLIGHTENMENT

The death of Louis XIV on September 1, 1715, closed an epoch; the date of 1715 is a useful starting point for the Enlightenment. But the beginnings of critical thought go back much further, to about 1680, where one can begin to discern a new intellectual climate of independent inquiry and the questioning of received ideas and traditions.

The earlier date permits the inclusion of two important precursors. Pierre Bayle, a Protestant forced into exile by the repressive policies of Louis XIV against the Huguenots, paved the way for later attacks upon the established church by his own onslaught upon Roman Catholic dogma and, beyond that, upon rationalist ideologies of all kinds. His skepticism was constructive, however, underlying a fervent advocacy of toleration based on respect for freedom of conscience. In particular, his *Dictionnaire historique et critique* (1697; 2nd ed., 1702) became an arsenal of knowledge and critical ideas alike for the 18th century.

Bayle and Fontenelle

Bayle's contemporary Fontenelle continued in Descartes's wake to make knowledge, especially of science, more accessible to the educated layman. His *Entretiens sur la pluralité des mondes* (1686; *A Plurality of Worlds*) explains the Copernican universe in simple terms, the author expounding his lessons with characteristic gallantry to an attractive marquise on six moonlit evenings in the park of her château. The *Histoire des oracles* (1687) complements this popular erudition by a rationalist critique of erroneous legends. Fontenelle helped to lay the basis for empirical observation as the proper approach to scientific truth.

Both Bayle and Fontenelle promoted the Enlightenment principle that the pursuit of verifiable knowledge was a central human activity. Bayle was concerned with the problem of evil, which seemed to him a mystery in which philosophical speculation was gratuitous and understandable by faith alone. But such unknowable matters did not at all invalidate the search for hard fact, as the *Dictionnaire* abundantly shows. Fontenelle, for his part, saw that the furtherance of truth depended upon the elimination of error, arising as it did from human laziness in unquestioningly accepting received ideas or from human love of mystery. Though both thinkers shared a pessimistic view of human nature, they also saw ways in which the human condition could be improved.

Montesquieu

The Baron de Montesquieu, the first of the great Enlightenment authors, demonstrated a liberal approach to the world fitting in with a pluralist view of society. His *Lettres persanes* (1721; *Persian Letters*) were at once successful and established his reputation. Purporting to be letters written to and by two Persians visiting France, they depict a contemporary Paris full of vitality and movement but precariously vulnerable to possible despotism.

His interest in social mechanisms and causation is pursued further in the *Considérations sur les causes de la grandeur des Romains et de leur décadence* (1734; *Reflections on the Causes of the Grandeur and Decline of the Romans*). To explain Rome's greatness and decline, Montesquieu invokes the notion of an *esprit général* ("general spirit"), a set of secondary causes underlying each society and determining its developments. Herein are the seeds of *De l'esprit des lois* (1748; *The Spirit of Laws*), the preparation of which took 14 years. This great work brought political discussion into the public arena in France by its insistence upon the need, in whatever form of society, to maintain liberty as prime object of concern.

Voltaire (François-Marie Arouet), on any count, bestrides the Enlightenment. Whether as dramatist, historian, reformer, poet, philosopher, or correspondent, for 60 years he remained an intellectual leader in France. A stay in England (1726–28) led to the *Lettres philosophiques* (1734), which, taking England as a polemical model of philosophical freedom, experimental use of reason, and respectful patronage of arts and science, offered a program for a whole civilization. In later years Voltaire's onslaught upon the power of the Roman Catholic Church became more direct, as he denounced its doctrines and practices in countless pamphlets and the *Dictionnaire philosophique* (1764; *The Philosophical Dictionary*), the vade mecum of Voltairean attitudes. He laboured on historical works all his life, producing most notably *Le Siècle de Louis XIV* (1751; *The Age of Louis XIV*) and the *Essai sur les moeurs* (1756; "Essay on Customs"), the latter a world history of a half-million words. Above all, it was the growth of civilizations and cultures that particularly commanded his attention and formidable energy. He is best remembered for the tale *Candide* (1759), a savage denunciation of metaphysical optimism ("all is for the best in the best of all possible worlds") that reveals a world of horrors and folly. In *Candide*'s characters the instinct of survival remains uppermost, however, and provides a ray of hope in an otherwise sombre scene. Candide at last renounces absolute truths as futile and settles for the simple life of "cultivating his garden." The *conte* ("tale") called *L'Ingénu* (1767) continued this lesson; Voltaire turned from metaphysics to social satire upon the corrupt French government (set with prudence retrospectively in Louis XIV's reign). Reformist appeals to justice were the main focus of Voltaire's writings in his last 20 years, as he protested against such outrages as the executions, religiously motivated, of Jean Calas and the Chevalier de La Barre.

Diderot

Another universal genius of the age was Denis Diderot. He occupied a somewhat less exalted place, however, essentially because most of his greatest works were published only posthumously. But his encylopaedic range is undeniable. Theorist of the bourgeois drama, author of the greatest French antinovel of the century (*Jacques le fataliste,* 1796), and the first great French art critic (*Salons*), Diderot seized on the vision of a world materialistic and godless yet pulsating with energy and the unexpected. *Jacques le fataliste* captures the fluidity of a disconcerting universe where nothing is ever quite clear-cut or totally under control. Jacques represents it suitably by believing in fatalism yet acting with decisiveness when he wishes, just as if he possessed free will.

Diderot's interest in the plasticity of matter, where categories such as animal, vegetable, and mineral never seemed as distinct as conventional thought suggested, combined with an interest in biology, nowhere better exemplified than in *Le Rêve de d'Alembert* (written 1769, published 1830; *The Dream of d'Alembert*). This work is written in the characteristic form of a dialogue, allowing Diderot to range free with speculative questions rather than attempt firm answers.

In his own day Diderot was best known as editor of the *Encyclopédie,* a vast work in 17 folio volumes of text and 11 of illustrations. He and Jean Le Rond d'Alembert inaugurated the undertaking in 1751, and Diderot edited alone from 1758 until the final volume of plates appeared in 1772. A summation of knowledge rather than a radically polemical enterprise, the *Encyclopédie* is, however, the epitome of the Enlightenment, disseminating information to improve the human lot, reduce theological superstition, and, in Diderot's words from his key article "Encyclopédie," "change the common way of thinking."

DRAMA

Tragedy and the survival of Classical form. Classical tragedy survived into the 18th century, most notably in the theatre of Voltaire, which dominated the Comédie-Française from the premiere of *Oedipe* (1718) to that of *Agathocle* (1779). But even in Voltaire a profound change in sensibility is apparent as pity reigns supreme, to the exclusion of terror. Tragedy, in the view of Fontenelle or the Abbé Dubos, should teach men virtue and humanity. Voltaire's *Zaïre* (1732) does just that, through the spectacle of Christian intolerance overwhelming the eponymous heroine, torn as she is between the religion of her French Catholic forefathers and the Muslim faith of her future husband, a Turk. No fatality of character destroys her, but simply the failings of Christians unworthy of their creed, allied to gratuitous and avoidable chance. Such pathos, often touching, is not the stuff of great tragedy.

Theatre of Voltaire

Marivaux and Beaumarchais. The best of 18th-century drama takes a different course. Pierre Marivaux wrote more than 30 comedies, mostly between 1720 and 1740, bearing for the most part upon the psychology of love. Typically, the Marivaudian protagonist is a refined young lady who finds herself, to her bewilderment or even despair, falling in love despite herself, thereby losing her autonomy of judgment and action. *La Surprise de l'amour,* a title Marivaux used twice (1722, 1727), becomes a regular motif, the interest of each play resting in the delicate changes of attitude and circumstance rung by the dramatist. His sympathy for the generally likable heroes and heroines stops short, however, of indulgence. The action is dramatic essentially because the characters' stubborn pride, central to their being, has to succumb to the demands of their instincts. Vanity, in Marivaux's view, is endemic to human nature. In *Le Jeu de l'amour et du hasard* (1730; *The Game of Love and Chance*), the plot of which is based on masters disguised as servants and vice versa, a heroine named Silvia experiences profound consternation at the quite unacceptable prospect of falling for a valet. When she learns the happy truth, her relief immediately gives way to a determination to force her lover Dorante into surrender while he still thinks her a servant. As her father puts it, "What an insatiable vanity of amour propre!" Many plays deal more explicitly with social barriers created by rank or money, like *La Double Inconstance* (1723) and *Les Fausses Confidences* (1737). As the subtlety of Marivaux's perceptions has become better understood, he has come to be regarded as the fourth great classic (after Corneille, Racine, and Molière) of the French theatre.

Pierre-Augustin Caron de Beaumarchais is best remembered for two comic masterpieces, *Le Barbier de Séville* (1775; *The Barber of Seville*) and *Le Mariage de Figaro* (1784; *The Marriage of Figaro*). Both are dominated by Figaro, a scheming dynamo of wit and generosity. He is a wholly free man in the first play, plotting his master Almaviva's conquest of Rosine, but he is obliged in the second play to defend his fiancée, Suzanne, against Almaviva's irresponsible meddling. (Some critics have detected a prerevolutionary quality to *Le Mariage,* but the evidence is too insubstantial to sustain this thesis.) As much as the sharpness of wit and character, the brilliance of structure wins admiration. All is movement and vicissitude, *Le Mariage* in particular with its 92 scenes (about thrice the average number in a Classical play) and profusion of theatrical "business" rising to the magisterial imbroglio of the final act.

Bourgeois drama. Yet Beaumarchais himself espoused the *drame bourgeois* ("bourgeois drama," or "middle-class tragedy") in his *Essai sur le genre dramatique sérieux* (1767). He wrote several *drames,* among them the sequel to *Le Mariage* in *L'Autre Tartuffe, ou La Mère coupable* (1792). The growing importance of sentiment on the stage had proved as inimical to Classical comedy as to Classical tragedy. More popular was a type of comedy both serious and moralistic, like *Le Glorieux* (1732; *The Conceited Count*) by Philippe Néricault Destouches, who claimed in the preface that he wanted to "purify society," or the *comédies larmoyantes* ("tearful comedies") of Pierre-Claude Nivelle de La Chaussée, which enjoyed great popularity when they appeared in the 1730s and '40s. Diderot's *Entretiens sur le fils naturel* (1757) gave a theoretical underpinning to the new mood. The author called for a "tragédie domestique et bourgeoise," realistic and affecting, able to inspire strong emotions and incline audiences to more elevated states of mind. This "genre sérieux" ("serious genre"), reacting against the articulate tirades of Classical tragedy, would draw on pantomime and tableaux or inarticulate speech rather than on eloquent discursiveness. Beaumarchais's own prescriptions run on similar lines. Though Diderot's plays did not live up to his theories, the emphasis upon middle-class virtuousness was to be made dramatically effective in Michel-Jean Sedaine's *Le Philosophe sans le savoir* (1765; "The Unwitting Philosopher"; Eng. trans., *The Duel*). Louis-Sébastien Mercier transposed the same formula a class downward into the artisan world in *La Brouette du vinaigrier* (1775; "The Barrel-load of the Vinegar Merchant"). Both achieved greater fidelity of representation on the stage. But the success of the *drame bourgeois* was short-lived, perhaps because it attempted the incompatible aims of being true to life and inculcating idealistic attitudes.

POETRY

The emphasis upon reason, science, and philosophy may explain the absence of great poetry in the 18th century. The best verse is that of Voltaire, whose chief claim to renown during most of his lifetime was as a poet. In epic, mock-epic, philosophical poems, or witty society pieces he was preeminent; but to the modern critic the lyrical effusion or linguistic intensity that might indicate genius seem to be missing.

THE NOVEL

The success of the novel is a more positive story. Despite official opposition and occasional censorship, the new genre developed apace. The first great 18th-century exemplar is now seen to be Robert Challes, whose *Illustres françaises* (1713), a collection of seven tales intertwined, commands ever greater attention for its serious realism and disabused candour anticipating Stendhal. As the bourgeois spirit acquired a more prominent place in society, the *roman de moeurs* became important, most notably in the novels of Alain-René Lesage: *Le Diable boiteux* (1707; *The Devil upon Two Sticks*) and more especially *L'Histoire de Gil Blas de Santillane* (1715–35; *The Adventures of Gil Blas of Santillane*). The latter, a loose-knit picaresque novel, recounts its hero's rise in society and concomitant moral education, set against a comprehensive picture of the surrounding world. Characterization and sensibility receive greater attention in the novels of the Abbé Prévost, whose reputation has become more broadly based with the publication (beginning in 1969) of his complete works. He is best known, however, as the author of the *Histoire du chevalier des Grieux et de Manon Lescaut* (1731), an ambiguous mixture of disinterested passion and shabby criminality in which des Grieux, a young scapegrace but also a man of the most exquisite sentiments, sacrifices himself to the amoral, delicate, and forever enigmatic Manon. In this tragic tale love conquers all, but it constantly needs vulgar money to sustain it. Tears and swoonings abound, as do precise notations of financial costs. This blend of traditional romance and sordid realism, never quite one or the other, combines with the ambivalent characterization of the chevalier to create a masterpiece.

By contrast, Marivaux as novelist devoted his main energies to psychological analysis and the moral life of his characters. His two great narratives, *La Vie de Marianne* (1731–41) and *Le Paysan parvenu* (1734–35; "The Fortunate Peasant"), follow one single character recounting, as in *Manon Lescaut,* his past experience. But unlike the tone of *Manon Lescaut,* the comic note prevails in Marivaux's novels as Marianne and Jacob make their way upward in society. Reflection upon conduct becomes more important than conduct itself; the narrator, now of mature years, comments and endlessly interprets his actions when young and still in transit socially. The result provides a rich density of feelings, meticulously analyzed or finely suggested. Both protagonists are morally equivocal, born survivors with an eye for the main chance, yet they also are attractive because they reveal themselves so disarmingly and because they are capable of disinterested and honourable actions. What abides, however, is the portrait of one particular consciousness, unceasingly informative on the mysteries of the human mind and heart.

Rousseau

The preeminent name associated with the sensibility of the age is that of Jean-Jacques Rousseau. He promoted the cult of nature, lakes, mountains, and gardens, in contrast to the false glitter of cultured society. He called for a new way of life attentive above all to man's innate sense of pity rather than dependent upon meretricious reasoning; espoused untutored simplicity and the true equality of all because all men share a capacity for feeling; and aimed at total sincerity in his confessional writings and thereby gave birth to the modern autobiography. Thus he stands as one of the greatest thinkers of his time, alongside, and generally in opposition to, Voltaire. He established the modern novel of sensibility with the resounding success of his *Julie: ou La Nouvelle Héloïse* (1761), a novel about an impossible, doomed love. He composed an indispensable classic of educational theory with *Émile: ou De l'éducation* (1762), which traces the program of an ideal education from birth to marriage. The hero, brought up away from corrupting society, becomes a truly moral person, in keeping with the principles of natural man. Rousseau stresses the importance of feeling and spontaneity of action over purely theoretical doctrine; religious sensibility is an essential element of Émile's makeup.

The sharp hostility toward contemporary society already evident in his *Discours sur les sciences et les arts* (1750) is more profoundly elaborated in the *Discours sur l'origine et les fondements de l'inégalité parmi les hommes,* sometimes called the *Second Discours* (1755). Social inequality has come about because men have allowed their God-given right of freedom to be usurped. Our elegant, civilized society is a sham, whose reality is endless posturing, hostility, injustice, enslavement, and alienation. The possibly revolutionary implications are to some extent spelled out in the *Contrat social* (1762; *A Treatise on the Social Contract*). Liberty and equality can be reestablished, according to Rousseau, by a new social pact of all with all, willingly accepted, obeying the *volonté générale* ("general will"), which alone has total sovereignty. In this way passive subjects of the state will be turned into active citizens zealous for the public good. But this act will require a moral transmutation, whereby men use their reason properly, in the exercise of selfless virtue; this emphasis on reason is underlined by the highly abstract character of the work. Commentators have differed widely in their readings of the *Contrat social* as a liberal or totalitarian document, but Rousseau at least saw himself as unambiguously defending freedom.

The *Confessions*

Rousseau's struggle toward a morality based on transparent honesty is continued in the *Confessions* (written 1765–70), where he seeks self-knowledge through awareness of the unconscious, the importance of childhood in shaping the man, and the role of sexuality; in so doing he anticipates modern psychoanalysis. This original exploration of the self is developed further in the *Rêveries du promeneur solitaire* (written 1776–78; *Reveries of a Solitary Walker*), which has been seen as foreshadowing even more strongly the Romantic movement and the literature of introspection of the next century.

The later 18th-century novel is dominated by the figure of Pierre Choderlos de Laclos and his masterpiece *Les Liaisons dangereuses* (1782; *Dangerous Acquaintances*), which gives a skillful account of erotic psychology as the libertine Valmont and his accomplice Mme de Merteuil plot the downfall of their victims. The tragic development is set in a Parisian society illustrating Rousseau's strictures: naturalness has given way to conformist expediency and cynicism, and the opportunity for spontaneous love has been lost. Laclos's novel delineates the void left in this heartless world; yet he also praises intelligence, which the highly crafted construction of the work brilliantly exemplifies. The detachment afforded by the structure of this epistolary novel precludes an assured understanding of the authorial viewpoint, and it has proved possible to interpret the work variously as an appeal to sensibility or as an exercise in pure cynicism. The ambiguity abides, preserving the novel's place among the classics of the genre.

The Marquis de Sade

By contrast, Jacques-Henri Bernardin de Saint-Pierre's *Paul et Virginie* (1788) often seems over-sentimental to modern tastes. It remains nonetheless a rich evocation of exotic nature in the tropical setting of Mauritius. Nicolas-Edme Restif (called Restif de la Bretonne) wrote at great length about the Parisian society, often criminal and equivocal, in which he moved. Given to verbosity, he yet evoked vividly the low world around him. More disturbing is the Marquis de Sade, whose search for maximum intensity of physical sensation, particularly in deriving pleasure through the infliction of pain, gave rise to the word sadism. According to Sade, because nature is engaged in destruction as in procreation, murder is natural and morally acceptable. The true libertine must replace soft sentiment by an energy aspiring to total freedom. Sade's insistence upon experimental rationalism owes much to the Enlightenment, of which his antisocial egoism is, however, only a perverted expression. But in works like *Justine, ou les malheurs de la vertu* (1791; *Justine, or The Misfortunes of Virtue*) he made the reader aware as never before that the search for fulfillment via the enjoyment of cruelty forms part of the human psyche. (H.T.M.)

Entering the 19th century: from the Revolution to 1850

REVOLUTION AND EMPIRE

The French Revolution of 1789 provided no clean break with the literary traditions of the Enlightenment. Many ways of thinking and feeling and most literary forms persisted with little change from 1789 to 1815. Indeed, the Napoleonic regime encouraged a return to the Classical mode. The insistence on formal qualities, notions of good taste, rules, and appeals to authority implicitly underlined the regime's centralizing, authoritarian, and imperial aims. This Classicism, or, strictly speaking, Neoclassicism, represented the etiolated survival of the high style and literary forms that had dominated the "serious" literature—and drama in particular—in France for almost two centuries. It is the complex series of reactions against this dominance, the search for new forms, the expansion of the vocabulary of literature, and the development of new criteria of taste that, together with the profound impact of the social and political events of the times, provide the main thrust of French Romanticism.

The poetry of André Chénier. André Chénier was executed during the last days of the Terror. His work first appeared in volume in 1819 and is thus associated with the first generation of French Romantic poets, who saw in him a symbol of persecuted genius. Although deeply imbued with the Classical spirit, especially that of Greece, he nevertheless exploited Classical myths for modern purposes. He began work on what he planned to be a great epic poem, "Hermès," a history of the universe and human progress. The completed fragments reflect the Enlightenment spirit but also anticipate the episodic epic poems of the Romantics (such as Victor Hugo's *Légende des siècles,* 1859–83). Chénier, though a moderate in revolutionary terms, was deeply committed in his politics. This is evident in the scathing fierceness of his *Ïambes,* many of which were written from prison shortly before his execution. His best known poems, however, are elegies that sing of captivity, death, and dreams of youth and lost happiness.

Revolutionary oratory and polemic. The intensity of political debate in Paris during the Revolution, whether in clubs, in the National Assembly, or before tribunals, threw into prominence the arts of oratory. Speaking in the name of reason, virtue, and liberty, and using the Roman republic or the city-states of Greece as frame of reference, revolutionary leaders such as Honoré-Gabriel Riqueti, comte de Mirabeau, infused the intellectual preoccupations of the Enlightenment with a sense of drama and passion. This renewal of rhetoric is echoed in the enormously expanded political press. To some extent the proclamations and communiqués of Napoleon prolonged this revolutionary eloquence.

Atala and *René*

Chateaubriand. The French Revolution made an émigré of François-René, vicomte de Chateaubriand, and his first major work, the *Essai . . . sur les révolutions* (1797), is a complex and sometimes confused attempt to understand revolution in general, the French Revolution in particular, and the individual's relationship to these phenomena. Chateaubriand took as his model the stance of the 18th-century *philosophe,* but his *Génie du Christianisme* (1802; "The Genius of Christianity") caught a new mood of return to religious faith based on emotional appeals and underlines the aesthetic superiority of Christianity. The impact of this work was enormous. Within it were two short narratives, *Atala* and *René.*The latter quickly came to represent the *mal du siècle,* the essence of Romantic sensibility, presenting an insecure, solitary, and disorientated young hero not dissimilar to Goethe's Werther.

Behind all of Chateaubriand's works lies the sense of a break, caused by the French Revolution, in a stable, ordered existence. His *Mémoires d'outre-tombe* (1848–50; "Memoirs from Beyond the Tomb"), the masterpiece he worked on most of his adult life and intended for posthumous publication, uses the autobiographical format to meditate on the history of France, the passing of time, and the vanity of human desires. His lyrical and rhythmic prose, which earned him the title Enchanteur, left a deep impression on many Romantic writers.

Mme de Staël and the debate on literature. Mme de Staël (Anne-Louise-Germaine Necker, baronne de Staël-Holstein) was truly encyclopaedic in her interests. Her contribution to intellectual debate far exceeded any narrow definition of literature. Liberal and, after her offer of support was rebuffed, fiercely anti-Napoleon in politics, eclectic in philosophy, mixing rationalism and spiritualism, and determinedly internationalist in her feeling for literature, she moved most easily in a world of ideas, surrounding herself with the salon of intellectuals she founded at Coppet, Switzerland. Her two novels, *Delphine* (1802) and *Corinne* (1807; *Corinna*), focus on the woman of talent persecuted by society. Her two most influential works, *De la littérature* (1800; "On Literature") and *De l'Allemagne* (1810; *Germany*), expanded conceptions of literature, claiming, for example, that postrevolutionary society required a new literature. She explored the contrast, as she saw it, between the literature of the south (rational, Classical) and the literature of the north (emotional, Romantic), and she showed the interest of foreign writers like Shakespeare, Ossian, and above all the German Romantics.

Many of these ideas emerged from discussions with August Wilhelm von Schlegel, author of the widely read *Cours de littérature dramatique* (1809, 1814), and from meetings with and readings of the Germans Goethe and Schiller. The Genevan economist and writer Jean-Charles-Léonard Simonde de Sismondi reinforced many of Mme de Staël's points in his *De la littérature du midi de l'Europe* (1813). This cosmopolitan cultural relativism was an inflammatory attitude in the prevailing Neoclassical literary climate.

ROMANTICISM

The *mal du siècle*

French Romanticism cannot be reduced to a listing of themes. Although one can point to a remarkable number of works that concentrate on the isolated and misunderstood individual, evoke nature and natural forces, and emphasize a religious sensibility, these elements are not specific to the early 19th century. The writings of certain French authors, such as Jean-Jacques Rousseau, have been described as pre-Romantic, but in general Romanticism in France developed later than in Germany or Britain. Its particular flavour is a result of the impact of revolutionary turmoil and the Napoleonic odyssey on French writers' sensibilities. The terms *mal du siècle* (world weariness; literally "ills of the century") or *enfant du siècle* (literally "child of the century") symbolize their distress. The latter is picked up and exploited in Alfred de Musset's autobiographical *Confession d'un enfant du siècle* (1836). Most French Romantics, whether they adopted a liberal or conservative attitude or whether they wished to ignore the weight of history and politics, asserted that their century was sick, and all were acutely aware of being products of their time. There was thus a new preoccupation with time and change. Romantics often retained the encyclopaedic ambitions of their predecessors, but faith in any simple notion of progress was shaken. Some distinction can be made between the generation of 1820, whose members wrote, often from an aristocratic viewpoint, about exhaustion, emptiness, loss, and ennui, and the generation of 1830, whose members spoke of dynamism—though often in the form of frustrated dynamism.

Foreign influences. When the émigrés who had fled from the effects of the Revolution trickled back to France they brought with them some of the cultural colouring acquired abroad (mainly in Britain and Germany) and this partially explains the paradox of aristocratic and politically conservative writers fostering new approaches to literature. Mme de Staël, as a liberal exile under Napoleon, was an exception. Travel had broadened intellectual horizons and had opened up the European cultural hegemony of France to other worlds and other sensibilities. The influence of the British Lord Byron's poetry and of the Byronic legend was particularly strong. Byron provided a model of poetic sensibility, cynicism, and despair, and his death in the Greek war of independence reinforced the image of the noble and generous but doomed Romantic hero. Italy and Spain, too, exercised an influence, though, with the exception of Dante, it was not their literature that attracted so much as the models of violent and exotic emotions these countries offered, with a proliferation of gypsies, bandits, poisonings, and revenge tales. (C.Sm.)

The *cénacles*

The poetry of the Romantics. The new climate was especially evident in poetry. The salon of Charles Nodier became one of the first of the literary groups known as the *cénacles;* later groups were to centre on Charles-Augustin de Sainte-Beuve, who is remembered chiefly as a literary critic. In addition to the outstanding poets of the period were a host of minor talents, and the Romantic upheaval in literature opened the way for a variety of possible developments, from the gentle lyricism of Marceline Desbordes-Valmore to the frenetic extravagance of Petrus Borel. For a time, about 1830, there was a marked possibility that French Romantic poetry might veer toward radical politics and the socialism of utopian writers like Saint-Simon, rather than in the direction of *l'art pour l'art,* or art for art's sake. The popularity of the songs of Pierre-Jean de Béranger is a reminder of the existence of an alternative current, political and satirical, beside the intimate lyricism and aesthetic preoccupations of typical Romantic verse. (R.C.Bu.)

Lamartine. Alphonse de Lamartine made an enormous impact as a poet with his *Méditations poétiques* (1820; *Meditations*). Using a restricted Neoclassical vocabulary, and unadventurous in versification, he nevertheless succeeded in creating through the musicality of his verse and often through the evocation of vaporous landscapes a sense of great longings unfulfilled. This soft-centred elegiac tone is tempered by occasional deep despair and Byronic revolt. The *Harmonies poétiques et religieuses* (1830; "Poetic and Religious Harmonies"), with their religious emotion, reinforce the quest for serenity, which remains threatened by unease and disquiet. *Jocelyn* (1836; Eng. trans., *Jocelyn*) and *La Chute d'un ange* (1838; "The Fall of an Angel") are intermittently successful attempts at epic, which tempted many French Romantics. An undercurrent in Lamartine's poetry is the preoccupation with politics. His essential contribution to French poetry was made in the 1820s. His verse seemed to engage his whole being, in contrast to most 18th-century poets, who had seen their art as merely a polite accomplishment.

The early poetry of Hugo. It was also in the 1820s that the powerful and versatile genius of Victor Hugo emerged. In his first poems he was a supporter of the monarchy and the church. This conservative Roman Catholic legitimism is a common strand in the poetic generation of 1820, and the debt to Chateaubriand's *Génie du Christianisme* is evident. These early poems lack the mellifluous quality of Lamartine's *Méditations,* but by the time of the *Odes et ballades* (1826) there are already hints of the Hugoesque mixture of intimate poetry, speaking of family relationships and problems of the ego; a prophetic and visionary tone; and an eagerness to explore a wide range of poetic techniques. Hugo called *Les Orientales* (1829) a useless book of pure poetry. It can be linked with the *l'art pour l'art* movement (see below), concentrating on the exotic and the visual combined with verbal and formal inventiveness. Hugo published four important collections in the 1830s (*Les Feuilles d'automne,* 1831; *Les Chants du crépuscule,* 1835; *Les Voix intérieures,* 1837; *Les Rayons et les ombres,* 1840), in which poetry of nature and family life is interwoven with a sad, solitary, hesitant, but never quite despairing quest for true consciousness. The poetry moves from the personal to the visionary and the prophetic, prefiguring in the lyric mode the epic sweep of much of his later work.

Vigny's *Destinées*

Vigny. In contrast to Hugo's scope, the poetry of Alfred-Victor, comte de Vigny, was more limited and controlled. In common with Hugo and many other Romantic poets, however, he proposed the poet as prophet and seer. For Vigny the poet is essentially a dignified, moralizing philosopher, using the symbol as rational expression of his philosophy. Broadly pessimistic in tone, emphasizing suffering and noble stoicism, he multiplied victim figures in his work, with the poet-philosopher as quintessential victim. His *Destinées* (1864), composed between 1838 and his death in 1863, exemplify the high spiritual aspiration that represents one aspect of the Romantic ideal. The control and concentration of expression is in contrast to the verbal flood of much Romantic writing and foreshadows more modern attitudes to poetic composition.

Musset. The young, brilliantly gifted Alfred de Musset quickly established his reputation with his *Contes d'Espagne et d'Italie* (1830). His exuberant sense of humour, an uncommon trait among Romantic poets, led him to use extravagant Romantic effects and at the same time treat them ironically. Later, a trajectory from dandyism through debauchery to a sense of emptiness and futility, sustained only intermittently by the linking of suffering with love, resulted in a radical dislocation of the sense of self. The "Nuits" poems ("Nuit de mai," "Nuit de décembre," "Nuit d'août," "Nuit d'octobre," 1835–37) express this purifying power of suffering in verse of sustained sincerity, purged of all the early showiness.

Nerval. For a long while Gérard de Nerval was seen as the translator of German literature (notably his adaptation of Goethe's *Faust*) and as a charming minor Romantic. Later critics have seen as his real contribution to poetry the 12 sonnets of *Les Chimères,* composed between about 1844 and 1854, and the prose poems added to the spiritual odyssey *Aurélia* (1855). The dense symbolic allusiveness of these latter works is the poetic transcription of an anguished, mystical quest that draws on the most diverse religious myths and all manner of literary, historical, occult, and esoteric knowledge. They represent one of the peaks of achievement of that side of the Romantic movement that sought in the mystical a key to the spiritual reintegration of the divided postrevolutionary self. The prose poem and the use of symbol link up with the poetry of Charles Baudelaire and Stéphane Mallarmé.

Romantic theatre. Some critics have been tempted to call Romantic theatre in France a failure. Few plays from that time remain in the active repertory, though the theatre was perceived throughout the period to be the dominant literary form. Quarrels about the theatre provided some of the most celebrated battles of Romanticism against Classicism.

Hugo. The first performance of Hugo's *Hernani* (1830) was one such battle, and Romanticism won a symbolic victory. *Hernani* followed Stendhal's call in the pamphlets *Racine et Shakespeare* (1823, 1825) for theatre that would appeal to a contemporary public, Hugo's own major theoretical statement in the preface to his play *Cromwell* (1827), and the production in Paris of several Shakespearean and historical dramas. (In particular an English troupe playing Shakespeare had had a sustained and triumphal season in Paris in 1827.)

Hernani drew on popular melodrama for its effects, exploited the historical and geographical local colour of an imagined 16th-century Spain, and had a tragic hero with whom the young Romantics were able to identify. These elements are fused by Hugo's lyric poetry to produce a dramatic spectacle close to that of Romantic opera. *Ruy Blas* (1838), in similar vein, mixes poetry, comedy, and tragedy with strong antithetical effects to provide the mingling of dramatic genres, which Hugo saw as the essence of Romantic drama. The failure of *Les Burgraves* (1843) is commonly given as the beginning of the end of Romantic theatre. Its grandiose epic conception on stage seemed merely false, abstract, and melodramatic to its audiences.

Vigny. Where Hugo's verse dramas tended to the lyrical and the spectacular, Vigny's most famous play, *Chatterton* (1835), in its concentrated simplicity, has many analogies with Classical theatre. It is, however, a bourgeois drama of the sort called for by Diderot and focusses on the suicidal poet Thomas Chatterton as a symbolic figure of the idealistic poet misunderstood by a bourgeois society—a typical Romantic fate.

Musset. Musset did not have public performance primarily in mind when writing most of his plays, and yet, ironically, he is the one playwright of this period whose works have continued to be regularly performed. In the 1830s he wrote a series of short comedies and *proverbes*—almost charades—in which light-hearted fantasy and the delicate hesitations of young love, rather in the manner of Marivaux, are contrasted with ironic pieces expressing underlying disillusionment.

The larger scale *Lorenzaccio* (1834) is the one indisputable masterpiece of Romantic theatre. A drama set in Renaissance Florence but with clear links to post-1830 France is combined with a brilliant study of a once pure but now debauched hero almost paralyzed by doubt. The world of wasted youth and lost illusions is evoked in a prose that at times resembles lyric poetry. The showy historical colour and the bluster typical of Romantic melodrama are replaced here by a deep sense of tragic poetry that stands comparison with Shakespeare.

THE NOVEL FROM CONSTANT TO BALZAC

The novel was the most rapidly developing literary form in postrevolutionary France, its enormous range allowing authors to deal flexibly with the changing relationship of the individual to society. The Romantic undergrowth encouraged the flourishing of such subspecies as the Gothic novel and the terrifying or the fantastic tale—the latter influenced in many cases by the translation from German of the works of E.T.A. Hoffmann—works that, when they are not simply ridiculous, seem to be straining to provide a fictional equivalent for the subconscious or an intuition of the mystical. Balzac was greatly influenced by this "infra-literature" in his attempt to achieve a synthesis of the spiritual and the material.

Benjamin Constant's *Adolphe* (1816), presented as a fictional autobiography, belongs to an important strand in the tradition of the French novel, namely the novel of concentrated psychological analysis of an individual, which runs from the 17th century to the present day. In that sense *Adolphe* has about it a Classical intensity and simplicity of line. Its moral ambiguity, however, together with the hesitations of the hero and his weakness, is related to the contemporary sense of moral sickness. In spite of the difference of style, there is a clear link with the themes of Chateaubriand's *René* and Étienne Pivert de Senancour's *Oberman* (1804).

The historical novel. The acute consciousness of a changed world after the Revolution and hence of difference between historical periods led novelists to re-create the specificity of the past, with a distinct preference for the Middle Ages and the Renaissance. Until about 1820 the Middle Ages was often regarded as a period of barbarism between Classical antiquity and the Neoclassical 17th and 18th centuries. But Chateaubriand's lyrical evocation of Gothic ruins and young royalist writers' attraction to a certain vision of feudalism provided a different evaluation of the period. The vogue for historical novels was at its strongest in the 1820s and was given impetus by the immense influence of the French translations of Sir Walter Scott. The best example of the picturesque historical novel is Hugo's *Notre-Dame de Paris* (1831; *The Hunchback of Notre Dame*). In it Hugo re-created an atmosphere of vivid, colourful, and intense 15th-century life, associating with it a plea for the preservation of Gothic architecture.

Influence of Sir Walter Scott's novels

A deeper reading of Scott's novels is implicit in some of Honoré de Balzac's works. Balzac not only evoked the surface or the atmosphere of a precise period but also examined the processes of historical transformation. Scott's studies of the aftereffects of the Jacobite rising can be paralleled by Balzac's analysis of the Breton counterrevolution in *Les Chouans* (1829). The historical novel ultimately became the staple of the popular novel, as in *Les Trois Mousquetaires* (1844; *The Three Musketeers*) by Alexandre Dumas *père*.

Stendhal. The works of Stendhal (Henri Beyle) constantly pose the questions "who am I?" and "where lies true happiness?" and often represent an individual attempt to define an aesthetic. His autobiographical sketches, such as *Vie de Henry Brulard* and *Souvenirs d'égotisme* (published posthumously in 1890 and 1892, respectively) give a fascinating insight into a highly critical intelligence trying to organize his experience into a rational philosophy while remaining aware that the claims of emotion will often undermine whatever system he creates. In many ways he is an 18th-century rationalist with a 19th-century sensibility.

He came to the novel form relatively late in life. *Le Rouge et le noir* (1830; *The Red and the Black*) and *La Chartreuse de Parme* (1839; *The Charterhouse of Parma*) are his masterpieces. Both present a young hero grappling with a decidedly nonheroic social and political environment. *Le Rouge et le noir* focusses on France in the late 1820s, whereas *La Chartreuse de Parme*, situated in Stendhal's beloved Italy (where he lived for much of his adult life), often reflects a vision of the Italy of the Renaissance as much as that of the 19th century. In both works Stendhal is deeply concerned with happiness, the nature of the self, and the relationship of the individual to the social and the political. His quicksilver style is capable of embracing in rapid succession different emotions, ideas, and points of view, thereby creating a sense of spontaneity and making him one of the most individual of novelists.

George Sand. George Sand (Amandine-Aurore-Lucile Dudevant) was a dominant figure in the literary life of the 19th century. Her works are little concerned with the novel as an art form but are dominated by thematic or ideological considerations. For example, her semiautobiographical works study the iniquitous position of women in 19th-century society (*Indiana*, 1832; *Lélia*, 1833). From the later 1830s George Sand developed an interest in humanitarian socialism, an idealism tinged with mysticism, reflected in works like *Le Compagnon du tour de France* (1841) and *Consuelo* (1842). For a long while, however, her literary reputation rested on works such as *La Mare au diable* (1846), *François le Champi* (1848), and *La petite Fadette* (1849), sentimental stories of country life tinged with realistic elements. She is an excellent example of the sentimental socialists involved in the 1848 Revolution.

Nodier, Mérimée, and the conte. Charles Nodier and Prosper Mérimée both exploit the short story and the novella. Nodier specialized in the *conte fantastique* ("fantastic tale") to explore dream worlds or various forms of madness, as in *La Fée aux miettes* (1832), and succeeded thereby in suggesting the importance of the role of the unconscious in human beliefs and conduct. Mérimée also used inexplicable phenomena, as in *La Vénus d'Ille* (1837), to hint at repressed aspects of the psyche or the irrational power of passion. More commonly, combining a Classical analytical style with Romantic themes, he directed a cool, ironical look at violent emotions. Short stories like *Mateo Falcone* (1829) and *Carmen* (1845) are peaks of this art.

Balzac, too, was an excellent short-story writer, often continuing in the Romantic mode the philosophical *conte* of the 18th century.

Balzac. Honoré de Balzac is perhaps best known for his *Comédie humaine,* the general title of a vast series of more than 90 novels and short stories published between 1829 and 1847. In these works he concentrated mainly on an examination of French society from the Revolution of 1789 to the eve of the 1848 Revolution, organically linking realistic observation and visionary intuition, while at the same time seeking to analyze the underlying principles of this constantly developing world. He ranged back and forth, often within the same novel, from the philosophical to the social, the economic, the legal; from Paris to the provinces; and from the summit of society to the *petit-bourgeois,* studying the destructive power of what he called thought or passion or vital energy. The search for the absolute or the unifying principle corresponds to one of the highest aspirations of the Romantic period. By using techniques such as the recurrence of characters in several novels, Balzac gives a temporal density and dynamism to his works. The frustrated ambitions of his young heroes (Rastignac in *Le Père Goriot,* 1835; Lucien de Rubempré in *Illusions perdues,* 1837–43) and the subjection of women, particularly in marriage, are used as eloquent markers of the impasse into which bourgeois liberalism had led the French Revolution. He emphasized the dissolving power of money and the every-man-for-himself individualism unleashed by the Revolution. In spite of this pessimistic view of society, however, Balzac remained fascinated by the power and energy of life in all its manifestations.

19TH-CENTURY THOUGHT

Literary criticism and journalism. The passionate, even virulent, political journalism of the revolutionary period soon slowed to a trickle under Napoleon. Literary debate interwoven with political considerations was renewed after 1815, and a shifting spectrum of royalist Romantics and Classical liberals moved toward a Liberal-Romantic consensus about 1830. The young critic Sainte-Beuve, himself the author of poems, supported Romanticism about 1830, though progressively detaching himself from it as he elaborated his biographical critical method. Criticism in the major literary reviews tended to be from a modified Classical viewpoint throughout the 1830s and even the 1840s, the Romantics replying in inflammatory prefaces to their works. The surge in newspaper circulation after 1836 tended to create a more "popular" market for serialized novels with strong melodramatic effects, as in Eugène Sue's *Mystères de Paris* (1842–43; *Mysteries of Paris*).

Sainte-Beuve

Historical writing. Early 19th-century historians were committed to historical erudition, but their works often seem closer to the world of literature. Augustin Thierry's narratives present the histories of England and France in terms of race (Normans versus Saxons and Franks versus Gallo-Romans). This is essentially a poetic concept close to that of Sir Walter Scott's *Ivanhoe.* Similarly, the early volumes of Jules Michelet's *Histoire de France* (1833–44) are based on the poetic idea of intuitive sympathy leading to a total resurrection of the essence of a past period encapsulated in imaginative symbolic figures. Alexis de Tocqueville represents a turning away from Romantic historiography in his great analytical studies of social principles in *De la démocratie en Amérique* (1835–40; *Democracy in America*) and *L'Ancien Régime et la Révolution* (1856; *The Old Regime and the Revolution*).

The intellectual climate before 1848. The early 19th century saw the renewal of interest in religion, ranging from the sentimental religiosity of Chateaubriand to the traditionalist theology of Louis-Gabriel-Ambroise, vicomte de Bonald and Joseph de Maistre, but 18th-century sensualism continued and was developed by the Idéologues, proponents of the Idéologie movement. Claude-Henri de Rouvroy, comte de Saint-Simon, and his followers tried to evolve a synthesis, which proved unstable, between socialistic scientific analysis, particularly of economics, and Christian belief. Félicité Lamennais, a Roman Catholic priest, moved toward a Christian socialism that ultimately estranged him from the church. The whole of the first half of the century is marked by such idealistic attempts to reconcile the head and the heart in imaginative syntheses.

Renan, Taine, and the movement of ideas in the second half of the 19th century. After the failure of what was seen as the woolly idealism of the 1848 Revolution a consciously scientific spirit came to dominate the study of social and intellectual life. Auguste Comte's *Cours de philosophie positive* (1830–42) fathered this new school of thought, called Positivism. It became almost a new religion. Ernest Renan adapted this scientific approach to the study of religion itself, notoriously in his *Vie de Jésus* (1863; *Life of Jesus*). Hippolyte Taine continued this Positivist analysis and emphasized the importance of biological science so as to produce a form of biological determinism to explain human conduct. His impact on the Naturalist literary theories of Émile Zola was crucial. (C.Sm.)

Completing the 19th century: from 1850 to 1900

Literature in the second half of the 19th century continued a natural expansion of trends already established in the first half. Intellectuals and artists remained acutely aware of the same essential problems: the nature of man, his relationship with the universe, the guarantees of morality, the pursuit of beauty, and the duties of the artist. But as writers became progressively alienated from the official culture of the Second Empire (1852–70), the forms of their revolt became more and more disparate. While the principles of Positivism were easily assimilated to the materialist pragmatism of the developing capitalist society, even many rationalist thinkers were drawn to forms of Idealism that placed faith in progress through science; and the antirationalist and antiutilitarian writers diverged into various types of mysticism and aesthetic formalism.

NEW DIRECTIONS IN POETRY

Gautier and l'art pour l'art. The greatest changes occurred in poetry. By the end of the 1830s Romantic poetry had narrowed in range, becoming synonymous with the direct expression of personal feelings and ideals in alexandrine verse paragraphs. Turning his back on his own earlier attempts to treat grand themes in the grand manner, Théophile Gautier sought a new direction for lyric poetry by linking idealism with aesthetics. From the first edition of *Émaux et camées* (1852; "Enamels and Cameos") to the posthumously published *Derniers vers* (1872), he devoted himself to a form of literary miniature painting, attempting to make something aesthetically valid out of subjects for the most part deliberately chosen for their triviality. The fashion for linking poetry with the plastic arts had grown up during the 1840s. Gautier simply developed the implications of this trend to the ultimate, concentrating on the language of shape, colour, and texture and limiting form almost exclusively to the very restrictive octosyllabic quatrain. Even themes that in his prose fiction suggest a genuine spiritual unrest, such as the fluid nature of identity or the destructive power of love, become the occasion for virtuoso ornamental elaboration. Many of the poems are stylized, sometimes ironic, treatments of amatory themes; others play with images of everyday life; but the best are transpositions from one art form to another, particularly those based on music ("Symphonie en blanc majeur," "Variations sur le Carnaval de Venise," and "Contralto").

Leconte de Lisle and Parnassianism. Gautier's cult of form is also to be met in the work of Théodore de Banville, notably in *Les Stalactites* (1846) and *Les Odes* *funambulesques* (1857). But the reaction against the expression of personal emotion in rambling rhetorical verse was not confined to the formalism of the *l'art pour l'art* poets. Charles-Marie-René Leconte de Lisle, who came to be labelled the founder of Parnassianism, took a different approach in his *Poèmes antiques* (1852), *Poèmes barbares* (1862), and *Poèmes tragiques* (1884). Although his theoretical pronouncements about the supremacy of beauty suggest affinities with Gautier, Leconte de Lisle was far from believing that the subject matter of poetry was of no significance. He wanted his poetry to transmute knowledge into a higher form of truth, and he believed in the necessity of systematic research prior to composition. The highly material surface of his poems is used to disguise deeply felt nihilistic metaphysical beliefs. For Leconte de Lisle the history of mankind presented a long, slow decline from the golden age of antiquity, leading inevitably toward the cosmic annihilation that post-Darwinian biologists saw as the natural end of evolution. The stories recounted from European and Oriental mythology, and the portraits of exotic animals and landscapes, though superficially scientific in their blending of scholarly documentation and objective narrative manner, all distill the same sense of revolt against a destiny that binds mankind to expiate crimes it is fated to commit. The style of the poems is not in fact concerned with the transcription of surfaces: the physical details are used to create moods that form symbolic analogues for the philosophical ideas.

Leconte de Lisle's ostensible manner and matter were taken up with enthusiasm by younger contemporaries. But only *Les Trophées* (Eng. trans., *Les Trophées;* "The Trophies"), the exquisitely miniaturist sonnets of José Maria de Heredia, written over a quarter of a century but not published until 1893, are still read, and even these have an appeal more like that of Meissen china fruit than the emotional or intellectual force normally associated with poetry.

Baudelaire. It is significant that Gautier, Hugo, and Leconte de Lisle were the three contemporary French poets for whom Charles Baudelaire felt the greatest admiration, although he had no time for formalism, didacticism, or the cult of antiquity. Antithetical in all things, Baudelaire was torn between the desire to express a metaphysical anguish more urgent and subjective than that of the Romantics and an aesthetic conviction that the effectiveness of art depended on precision and control. It is as misguided to look for consistency in Baudelaire's critical works (*L'Art romantique, Curiosités esthétiques,* both published posthumously in 1868) as in his poetry, since his ideas evolved constantly and in some cases radically throughout his most creative period (1845–64). To two basic ideas, however, he remained constant: that only the artist can create meaning out of the raw material of life, and that the material world is irredeemably corrupted by original sin. The first of these is responsible for the importance which he assigns to intuition, imagination, synesthesia, and the necessity for the artist to plunge himself into the world about him. The second led him to an impasse: the artist could only rise above material corruption through the creative act, but the creative act could not occur without the stimulus of corrupt reality. Baudelaire was accordingly a poet deeply concerned with the relationship between morality and art, which he located in the effective transposition of the artist's special perceptions of the world, and he was genuinely distressed by the official condemnation of the first edition of *Les Fleurs du mal* (1857; *The Flowers of Evil*) on a charge of obscenity provoked by its supposed erotic realism.

Les Fleurs du mal

The tensions within Baudelaire reached their height in the second edition of *Les Fleurs du mal* (1861). The collection is loosely structured to present a "self" who struggles to transcend the limitations of the material world. The struggle is presented in a series of experiences that start with love (mostly physical), move out into the ugly urban environment of contemporary Paris, and gradually descend through increasingly vicious experiences until only death offers the possibility of new stimulus sufficient to keep the creative consciousness of the poet-hero alive. The stylistic antitheses mirror the content. Within individual poems Baudelaire shifts between the rhetorical, the impressionist, the abstract, and the intensely physical. He balances banality and originality, the prosaic and the melodic, to emphasize the eternal interdependence of opposites, which he sees as the essence of man's condition.

In the last years of his life Baudelaire tried to extend the literary means at his disposal by experimenting with prose poetry. The range of themes in the posthumously edited *Petits poèmes en prose* is similar to that of *Les Fleurs du mal,* though the balance is different: urban landscapes, the ambiguous relationship of artist and crowd, and the degradations of poverty are given more space than is love. The relative freedom of the prose form allowed more shifts of tone; the juxtaposition of the ironic and the lyrical; and the interweaving of anecdote, narrative, and reflection. The best poems, such as "La Chambre double," make a positive use of this new freedom.

The later poetry of Victor Hugo. It comes as a shock to realize that though the second half of the 19th century is habitually treated as a period of reaction against Romanticism, nearly all the major poetry of Hugo was, in fact, published after 1850. The three collections *Les Châtiments* (1853), *Les Contemplations* (1856), and *La Légende des siècles* (1859, 1877, 1883) are linked by their epic quality. Different as they are in content, intention, and tone (though less varied in style), each is loosely structured to create an overall unity. *Les Châtiments,* written from exile in the Channel Islands and published clandestinely, is a hymn of hate against the mediocrity, callousness, and greed of Louis-Napoleon and the society of the Second Empire, interspersed with outbursts of compassion for the poor and oppressed. The poems are arranged so as to emphasize the darkness of the present and the light of the future, as Hugo proclaims his optimistic belief in the eventual triumph of peace, liberty, and social justice. Though the most monotonous of the three great works in content, it is the most varied in style, as if to offset the repetitiousness of the invective. In contrast to this political saga, *Les Contemplations* embodies Hugo's philosophical attitudes. It presents the poet as seer, penetrating the mysteries of creation and recounting the metaphysical truths perceived. The first four books contain much nature poetry, elegy, and love lyrics. The portentous and obscure rhetorical enumerations of the later visionary sections have dated beyond recall. *La Légende des siècles* reveals the same urge to play the prophet. The poems are a series of historical and mythological narratives, borrowing some of the scientific spirit that informed Leconte de Lisle's work but with none of the same attention to preliminary scholarly research. Together they form an account of the history of human existence as Hugo saw it.

After the three epic cycles Hugo returned to writing short lyrics on personal themes (*Les Chansons des rues et des bois* [1865; "Songs of the Streets and the Forests"]; *L'Art d'être grand-père* [1877; "The Art of Being a Grandfather"]), although he never abandoned his role as didactic poet, as the collections churned out in the 1880s testify.

REALISM IN THE NOVEL

Diversity among the Realist school. The label Realism came to be applied to literature by way of painting as a result of the controversy surrounding the work of Gustave Courbet in the early 1850s. Courbet's realism consisted in the emotionally neutral presentation of a slice of life chosen for its ordinariness rather than for any intrinsic beauty. Literary realism, however, was a much less easily definable concept. Hence the loose use of the term in the late 1850s, when it was applied to works as various as Flaubert's *Madame Bovary,* Baudelaire's *Fleurs du mal,* and the social dramas of Alexandre Dumas *fils.* Even the members of the so-called Realist school were not entirely in agreement. Edmond Duranty, cofounder of the monthly journal *Réalisme* (1856), supported the view that novels should be written in a plain style about the ordinary lives of middle- or lower-class people, but he insisted that the Realists' main aim should be to serve a social purpose. Jules-François-Félix Husson (known as Champfleury), an art critic and novelist, stressed the need for careful research and documentation and rejected any element of moral

The brothers Goncourt

intention. The practice of those labelled Realists was even more diverse than their theory. The writers who most fully realized Champfleury's ideal of a documentary presentation of the day-to-day, Edmond and Jules Goncourt, were also the most concerned with that aesthetic perfection of style which Duranty and Champfleury rejected in practice as well as in principle. In the Goncourts' six jointly written novels that appeared in the 1860s, and in four further novels written by Edmond Goncourt after his brother's death, plot is reduced to a minimum and the interest of the novel is divided equally between stylistic bravura and the minutely documented portrayal of a milieu or a psychological state—the upbringing of a middle-class girl in *Renée Mauperin* (1864), or the degenerating life-style of a female servant in *Germinie Lacerteux* (1864; *Germinie*).

Gustave Flaubert. The problem with the Realists was that each had his own definition of reality and his own recipe for how to transcribe it. It is easy to see why Gustave Flaubert was so firm in dissociating himself from such writers as Champfleury and Duranty, given that his own work undermined all sense of stability in perceptions and values by emphasizing the idea that reality is relative to the person who perceives it. Furthermore, Flaubert rejected any idea of transposing a slice of life onto the page in everyday language. For him, only art could give meaning to the raw material provided by the external world, and only through art could language be lifted above the emptiness of its everyday function so as to record objectively the author's perception of the world.

Flaubert's juvenilia, of which the first version of *L'Éducation sentimentale* (1845) is the most important work, show the writer's struggle to control his own instinctive idealism and to find a way of reconciling his belief in the primacy of facts with his rejection of contemporary petty materialism. His fascination with escapism and Romantic excess was to reappear in *Salammbô* (1863; *Salambo*) and *La Tentation de Saint-Antoine* (1874; *The Temptation of Saint Anthony*), in which he portrayed exotic subjects in a heightened lyrical fashion. But his major novels, *Madame Bovary* (1857) and *L'Éducation sentimentale* (1870; *Sentimental Education*) show trivial lives frittered away in hopeless attempts to transcend the banality and dishonesty of the modern world. Emma Bovary destroys herself by taking empty abstractions—passion, happiness—as concrete absolutes, and by attempting to base her life on such concepts. In her efforts to make the world (and more specifically the men) around her fit her preconceived image, she ignores the nature of material reality itself, as symbolized by money, and is inexorably drawn onward to financial ruin and suicide. Emma's own mediocrity is matched by that of the provincial society in which she lives, and her illusory view is paralleled by the various illusions entertained by almost all the major characters. Most of these, however, like the apothecary Homais, realize that the way to succeed in life is to accept the official values promoted as absolute by the ruling bourgeoisie. *L'Éducation sentimentale* extends the study to cover the entire "generation of 1848," showing how all emotional, artistic, and social ideals are corroded by contact with reality. Its central character, Frédéric Moreau, is a passive version of Emma, and the symbolism of money gives way to that of prostitution—the sale of love, talent, and principle.

In his *Trois contes* (1877; *Three Tales*) Flaubert approached the same beliefs from the opposite angle, showing how the withdrawal from reality of the ascetic or saint, exemplified by Félicité, the simple servant in "Un Coeur simple," St. Julian in "La Légende de Saint-Julien l'Hospitalier," and John the Baptist in "Hérodias," when coupled with the metaphysical aspirations central to Christian belief, offers a form of supreme illusion that the material world cannot touch.

In all these works, and in his unfinished satirical novel, *Bouvard et Pécuchet* (published posthumously in 1881; *Bouvard and Pécuchet*), Flaubert returns consistently to the problem of communication. He was haunted by the way in which the conventional language of love, literature, journalism, politics, and science pretends to portray a fixed reality while in fact concealing the subjectivity of all perceptions. His obsessive reworking of every sentence that he wrote was motivated by the desire to create a style that would preserve the meaninglessness of "reality" while justifying by its own beauty the existence of the novels.

DRAMA AFTER 1850

Major writers of social drama

The society of the Second Empire, and indeed that of the early decades of the Third Republic, was not fond of seeing itself too accurately portrayed on the stage; yet at the same time, in reaction against the escapism and nonconformity of Romantic drama, its members wanted the stage to reflect contemporary values and preoccupations. Hence the predominance, from 1850 to 1890, of social drama on the one hand and light comedy, farce, and operetta on the other. Social drama, denied the use of political issues by censorship, confined itself to the tension between new money and old social position, the morality of financial speculation, and the threat to family life posed by extramarital sexual relationships—all themes touched upon previously in light comedy (*e.g.,* the plays of Eugène Scribe). The settings and character types related to the audience's milieu; hence the plays were considered to be realistic at the time, although their sentimentality, black-and-white morality, and melodramatic turns of plot make them seem highly artificial in modern terms. The major writers of social drama were Alexandre Dumas *fils* and Émile Augier, both of whom began by describing society and ended by prescribing for it. Dumas *fils* is best remembered for his romanticization of the courtesan in *La Dame aux camélias* (1848), the novel and play on which the libretto of Giuseppe Verdi's *La Traviata* was based, but the moralizing *Les Idées de Mme Aubray* (1867), with its plea for the social redemption of repentant fallen women, is more typical of his major works. Augier's morality was more solidly conservative than was Dumas's, as can be seen from one of his best known plays, *Le Mariage d'Olympe* (1855), which not only proposes that what makes a woman into a prostitute in the first place is an innate propensity to vice, but also suggests that if a fallen woman insinuates herself into a respectable family it is legitimate to shoot her. On the other hand, Augier's treatment of the venality of the press and the corruption of financiers in *Les Effrontés* (1861) is as trenchant and effective as comparable portraits in the Naturalist novelists.

Light comedy and farce similarly relied upon a thin layer of contemporary social relevance, with marriage, the *ménage à trois*, and the pretensions of the lower middle class as the main subjects. In farce, in particular, social criticism passed from being an end to a means. Although the characters found themselves in situations in which events followed a nightmare logic quite at odds with their own stereotyped realism and revealed the potential irrelevance of their vaunted rationality, the return to sanity at the end of the plays confirmed the audience's assumption that the world would ultimately always conform to expected and accepted standards. The classic examples of the genre are the plays of Eugène-Marin Labiche, notably *Le Chapeau de paille d'Italie* (1851; *The Italian Straw Hat*) and *Le Voyage de M. Perrichon* (1860).

The operetta

When their taste ventured into something more literary, Second Empire audiences were obliged to look to the fantastical comedies of Alfred de Musset, written 30 years earlier but not staged until the 1850s and '60s. In light comedy proper and costume drama the leading figure of the age was Bernard Shaw's bugbear, Victorien Sardou. But the most successful genre of all was undoubtedly operetta, in particular the absurd comedies of the collaborators Henri Meilhac and Ludovic Halévy. Main examples of their work, set to music by Offenbach, are *La Belle Hélène* (1865), *La Vie parisienne* (1866), and *La Grande-Duchesse de Gérolstein* (1867). It is a reflection on the society of the day that *La Belle Hélène,* in which a frivolous pastiche of classical legend is spiced by an acute satire on the manners, morals, and values of the court of Napoleon III, was the nearest thing to political satire that the French stage could boast for 20 years.

The Franco-German War and the consequent collapse of the empire had little perceptible effect on mainline theatre. Offenbach lost favour because of his German associations, but Augier, Dumas, and Sardou continued to flourish. Attempts by other writers (Flaubert, the Goncourts, Zola) to establish a more genuinely realistic form of theatre failed, partly because public taste and theatrical commercialism made experiment nearly impossible, and partly because the plays written were either theatrically unsatisfying or lost much of their realism in reworking novels for the stage. The only effective Naturalist dramatist was Henry-François Becque, whose *Les Corbeaux* (*The Vultures*), first performed in 1882, and *La Parisienne* (1885; *Parisienne*), without completely ridding themselves of the structural formulas of the well-made play, came nearer than any other drama of the period to portraying brief moments in ordinary lives. That Becque owed his success to André Antoine, the founder and director of the Théâtre Libre (1887–96), is symptomatic of the way in which literary theatre in the last decades of the century was largely dependent on small-scale directorial experimentation. Antoine aimed at creating a unity between the staging (decor and acting style) of a play and its content, in the interest of total realism. From 1891 Paul Fort, founder of the Théâtre d'Art, and his successor, Aurélian-François-Marie Lugné-Poe, who restyled the company as the Théâtre de l'Oeuvre, applied Antoine's principles to the creation of antinaturalistic theatre. It was these little experimental companies that principally staged Symbolist plays, notably the works of Maurice Maeterlinck, with their emphasis on the suggestion of forces beyond reality and of meaning which lies beyond the explicit surface of language. But the audiences for *L'Intruse, Les Aveugles,* and *Pelléas et Mélisande* were small and select, and the significance of such theatrical innovation only became felt more widely in the following century.

Experimental theatre

NATURALISM

The argument for the existence of a Naturalist school of writing depends on the joint publication, in 1880, of *Les Soirées de Médan,* a volume of short stories by Émile Zola, Guy de Maupassant, Joris-Karl Huysmans, Henry Céard, Léon Hennique, and Paul Alexis. But it is doubtful whether Naturalism can justifiably be regarded as more than an extension of the methods of interpreting reality already in vogue. The Naturalists purported to take a more scientifically analytical approach to the presentation of reality than had their predecessors, treating dissection as a prerequisite for description. Hence Zola's attachment to the term *naturalisme,* borrowed from Hippolyte Taine, the Positivist philosopher who claimed for literary criticism the status of a branch of psychology. Unfortunately, it is difficult to find a coherent exposé of the Naturalist theoretical position. Zola's work notes are understandably fragmentary, and his public statements about the novel are all distorted by their polemical purpose—particularly the notorious essay "Le Roman expérimental" (1880; "The Experimental Novel"), in which he developed the untenable parallel between the methods of the novelist and of the experimental scientist. An examination of the views held in common by Zola, Maupassant, and Huysmans indicates that the basis of Naturalism can best be defined as the careful study of a given environment, the application of a mechanistic theory of psychology, and the rejection of any sort of idealism or escapism. However, like Flaubert the Naturalists did not see reality as fixed: Zola and Maupassant accepted as inevitable the transposition of reality through the temperament of the individual writer.

Émile Zola. The common points of Naturalist practice are even more elusive than their theory. Zola's Naturalism depends on the extensive documentation that he undertook prior to the writing of each novel. This extensiveness is emphasized by the subtitle of his 20-novel cycle *Les Rougon-Macquart: Histoire naturelle et sociale d'une famille sous le second Empire* (*Natural and Social History of a Family Under the Second Empire*). The linking of so many novels through a single family and the emphasis on the deterministic effects of heredity and environment confirm the scientific purpose. Zola's canvas is broader than Flaubert's, or even Balzac's: he handles subjects as diverse as a miners' strike in *Germinal* (1885; Eng. trans., *Germinal*), working-class alcoholism in *L'Assommoir* (1877; Eng. trans., *L'Assommoir;* "The Drunkard"), the sexual decadence of the upper classes in *La Curée* (1872; *The Kill*) and *Nana* (1880; Eng. trans., *Nana*), and the ferocious attachment of the peasantry to their land in *La Terre* (1887; *Earth*). But there are countless examples of manipulation of facts, particularly in the chronology of the novels, which show that for Zola documentary accuracy was not paramount. Indeed, his work notes reveal that he saw the scientific principles underlying the novels as a literary device to hold them together and thus strengthen the personal vision of reality that they contained. The sense of period and family unity is soon submerged, as Zola becomes both poet and moralist in his portrayal of contemporary values. All the major novels are dominated by symbolically anthropomorphized forces that control and destroy both individual and mass. Thus the mine in *Germinal* is represented as a voracious beast devouring those who work in it. This tendency to symbolism can be seen in an even more extreme form in the reinterpretation of the Genesis story in *La Faute de l'abbé Mouret* (1875; "The Transgression of Abbé Mouret"; Eng. trans., *The Sinful Priest*). As the cycle progresses, the sense of a doomed society rushing toward the Apocalypse grows, and it is indeed confirmed in the penultimate novel, on the Franco-German War, *La Débâcle* (1892; *The Debacle*).

The trilogy *Les Trois Villes* ("The Three Cities") and the unfinished tetralogy *Les Quatres Évangiles* ("The Four Gospels"), which followed *Les Rougon-Macquart,* became progressively more didactic, laying bare the obsessions with scientific progress, socialist humanitarianism, and the rejection of Roman Catholicism, which had been present in a concealed form in the earlier novels. Zola's contribution to French life after *Les Rougon-Macquart* lay more in his spirited intervention in the Dreyfus Affair than in what Henry James justifiably called Zola's "nursery moralities."

Guy de Maupassant. Of the other Naturalists only Maupassant is widely read, though Céard's *Une Belle Journée* (1881) and the four early novels of Huysmans, particularly *A vau-l'eau* (1882; *Down Stream*), have been unjustly neglected. Maupassant had the advantage of being a protégé of Flaubert, but his style is markedly less detached than his master's. Many of his short stories, whether set in Normandy or Paris, rely on sharply reductive, satirical techniques directed against his favourite targets—women, the middle classes, the Prussians—and designed to bring out hypocrisy and dishonesty as the central forces in human life. His novels achieve a greater sense of detachment and are truer to the aim expressed in his essay "Le Roman" (1887; "The Novel") to "write the history of the heart, soul and mind in their normal state." There is a progression in manner and matter from *Une Vie* (1883; *A Woman's Life*), with its echoes of *Madame Bovary,* through the detached but destructive portrait of the worlds of journalism and finance in *Bel-Ami* (1885; Eng. trans., *Bel-Ami*) to powerful evocation of the crippling effects of jealousy in *Pierre et Jean* (1888; Eng. trans., *Pierre et Jean*). His mastery of the psychological novel is confirmed by his portrait of the effects of old age on a pair of lovers in *Fort comme la mort* (1889; *The Master Passion*), and his considerable skill in presenting broader social forces at work is evident in his analysis of a small spa in *Mont-Oriol* (1887), which is reminiscent of Zola's work.

Maupassant's novels

THE REACTION AGAINST REASON

In the last decades of the century, particularly from 1880 onward, the opposition intensified between those creative writers who built their ideals around the material world and those who rejected physical experience as meaningless in itself. Whereas Baudelaire and Flaubert incorporated elements of both attitudes into their writings, the poets and novelists who followed them tended to take one or the other line to an extreme: hence the emergence within a short time of movements as disparate as Naturalism, Decadence, Symbolism, and the Roman Catholic Revival.

The Decadents: Verlaine and Laforgue. Decadence was a movement primarily associated with poetry but whose psychological basis is well illustrated in Huysmans' novel *À rebours* (1884; *Against the Grain*) and the *Culte du moi* trilogy (1888–91) by Maurice Barrès. It derives from the same extreme deterministic philosophy as Naturalism and has much in common aesthetically with Impressionism in that it isolates subjectively perceived moments of meaningless physical experience. The impetus to decadent poetry came partly from the study of Baudelaire and partly from the work of Paul Verlaine. Though much of his early poetry imitated the work of Baudelaire and the Parnassians, in the *Fêtes galantes* (1869), a group of pastiches, and his major collection, *Romances sans paroles* (1974; "Songs Without Words"), Verlaine created the blend of musicality, physical atmospherics, and sense of psychological distortion that constituted his greatest poetic achievement. In so doing, he used *impair* (odd) metres, ambiguous syntax, and unusual collocations of abstract and concrete concepts in a way that radically advanced the technical range of French verse.

Features of Decadent poetry

In Verlaine's work two impressions predominate: that only the self is important, and that the function of poetry is to preserve moments of extreme sensation and unique impression. These were the features, together with the experiments in form, on which the younger generation of poets seized in the 1880s. Hence the founding of the review *Le Décadent* in 1886, whose title consecrated a label originally coined by hostile critics. The poetic movement found its best exponent in Jules Laforgue, who brought together a subjectivism and pessimism fed by his studies in contemporary German philosophy and a genius for harnessing effects of poetic contrast. His first two published collections, *Les Complaintes* (1885) and *L'Imitation de Notre-Dame la Lune* (1886), are a series of variations on the themes of the flight from life, woman, and ennui, each explored through a host of frequently recurring images (*e.g.,* the wind, Sundays, the stock comic figure Pierrot). The conscious intellectual antitheses of the first collection are reduced in the second and disappear entirely in the *Dernier vers* (1890), in which the orchestrated cycle of recurrent images embodies the meaninglessness of life rather than explains it. To match these tenuous thematic patterns Laforgue used a fluid verse form shaped by rhythmic patterns and assonance, the first important example of free verse in French poetry.

Symbolism: Rimbaud and Mallarmé. The distinction between Decadence and Symbolism is surprisingly slight, although it should ostensibly be a distinction between the acceptance of the idea that life is meaningless and a mystical belief in meanings that transcend reality. But the principal Symbolist poets were not in a conventional sense metaphysicians, and their search for transcendence derived from a subjectivity as intense as that of Verlaine and Laforgue.

The narrowness of the distinction is well illustrated by the case of Arthur Rimbaud. Rimbaud's poetic creed is generally taken to be contained in two letters of 1871, in which he prescribes for the poet the need to explore the inner self, reorder perceptions of existence, and thereby become a visionary. The fiercely ironic view of contemporary society that emerges from his early poems reveals in him an element of the political revolutionary. The dreamlike poem "Le Bateau ivre" ("The Drunken Boat") and the poems written during the period of his infatuation with Verlaine (the so-called "last poems") show a more far-reaching revolution in the interpretation of reality, using an atmospheric, dislocated style similar to that of Verlaine's *Romances sans paroles* but with images more surreal and formal experiments more radical. By 1872 Rimbaud had carried his aesthetic revolution still further in his prose poems *Illuminations* (probably written from 1872 onward, and published posthumously), in which moods of destruction, revolt, elation, liberation, and frustration are suggested through visions that have no logical coherence. Critics are divided as to whether or not Rimbaud's other cycle of prose poems, *Une Saison en enfer* (1873; *A Season in Hell*), predated the writing of the last *Illuminations.* In either case the work reveals Rimbaud's own doubts as to whether his poetic method could create a vision capable of affecting the world outside itself. His loss of confidence in the magical powers of language led to his abandoning literature completely, at or about the age of 19.

Stéphane Mallarmé brought a very different temperament and intellectual background to bear on the problem of the inadequacies of the material world. While both Rimbaud and Mallarmé believed in the special power of language, Rimbaud's vision was centred on life, Mallarmé's on death. His early poems, such as "L'Azur" ("Azure") and "Les Fenêtres" ("The Windows"), reveal a man haunted by the awareness of a transcendent reality that he cannot attain and that frightens him. An intellectual and spiritual crisis in 1867 led him to a final acceptance of death as the only inevitability in life. But while this fact negated his personal significance, he found in it a new value for the products of his mind. Only the world of literature, born of the writer's imagination and fixed in the creative medium of language, was unassailable by the forces of chance. A meaning for life could only be found, therefore, in art. Ironically, although until his death Mallarmé pursued the vision of writing his Grand Oeuvre, he in fact published very little: the *Poésies* (1887), a handful of prose poems and essays, an unfinished prose drama, and "Un Coup de dés jamais n'abolira le hasard" (1897; "A Throw of the Dice Will Never Abolish Chance").

Though Mallarmé's philosophical crisis is important as an explanation of why he came to write as he did, his place in the history of poetry depends more on the techniques he evolved in the pursuit of his ideals. In the effort to escape from the limitation of surface reality, he came more and more to depend on elaborate techniques of evocation and suggestion. And in his desire to assert a God-like creative role he exercised tight control over all the formal elements of his verse—the very opposite of Rimbaud's experiments in liberation. As early as *L'Après-midi d'un faune* (1876; "The Afternoon of a Faun"; later interpreted musically by Claude Debussy) he concentrated on multiplicity of meaning: even a reductive reading of the poem has to see it as simultaneously the dream evocation of the faun's erotic desires and a meditation upon the creative impulse at an abstract level. The later "Prose (pour des Esseintes)," "Plusieurs sonnets," and "Autres poëmes et sonnets" are studies in the possibilities of language, in which, as in music, recurrent images and antithetical patterns communicate on an emotive level. The curious "Un Coup de dés," with its use of the typography to reflect the theme of chance, suggests that ultimately Mallarmé, like Rimbaud, felt defeated. He realized that the poet in his pursuit of the absolute via literature could never possess complete mastery of expression, and therefore could never exclude the hazard of random inspiration. His art was thereby reduced to the status of totally subjective self-exploration. Perhaps Huysmans was right to present Mallarmé as the ultimate decadent writer in *À Rebours.*

By an irony of literary history Symbolism as a historical movement strictly postdates Mallarmé. It derived its name from an article by Jean Moréas, who produced the first manifesto of the movement in 1886. The 1880s and '90s continued to be a period of artistic ferment, which produced many charming poems but no major poets. Musicality, myth, mysticism, and melancholia are the hallmarks of nearly all the best verse of the period and of nearly all the worst. Among those whose works have deservedly survived in anthologies are Henri de Régnier and Francis Viélé-Griffin and the Belgian poets Georges Rodenbach and Émile Verhaeren.

THE NOVEL LATER IN THE CENTURY

The development of the novel in the 1880s reflected the same thirst for absolutes as the new movements in poetry. But neither Decadence (with the exception of *À Rebours*) nor Symbolism generated prose fiction of lasting significance. There was a recrudescence of the *conte fantastique,* which found its foremost exponent in Auguste, comte de Villiers de L'Isle-Adam, as in his *Contes cruels* (1883; *Cruel Tales*). But the major trends in the novel were connected with the revival of Roman Catholicism and the growth of nationalism in the aftermath of the Franco-German War. The religious spirit was sometimes aesthetic, as in Huysmans' *La Cathédrale* (1898), sometimes visionary, as in Léon Bloy's *Le Désespéré* (1886) and *La Femme pauvre* (1897; *The Woman Who Was Poor*), in which a hysterical attack on bourgeois society is combined with an equally hysterical espousal of the then-fashionable doctrine of vicarious suffering. But the combination of Roman Catholic teaching and right-wing politics in the novels of Paul Bourget, beginning with *Le Disciple* (1889; "The Disciple"), is more typical of the spirit of the times. The antidemocratic, antirepublican views of Bourget were similar to those found in nationalist writers, notably Maurice Barrès. Barrès moved from decadent self-absorption to an extreme form of historical determinism. He saw the meaning of the individual as defined by his part in a collective inherited unconscious, which itself was defined by race. His trilogy *Le Roman de l'énergie nationale,* particularly *Les Déracinés* (1897), is an important document for an understanding of the attitudes of the French right during the Dreyfus Affair and between the world wars.

Anatole France

The only novelist of note who stood outside all these trends, and yet was a typical offspring of the age that produced them, achieving the double distinction of winning the Nobel Prize for Literature for 1921 and being put on the Index, was Anatole France (pen name of Jacques-François-Anatole Thibault). France made his initial reputation as a literary critic and author of psychological novels, but he rapidly became the personification of the pessimism fashionable after Germany's victory over France in 1870, an attitude typically expressed in the detachedly ironic exposure of human weakness in *Les Opinions de M. Jérôme Coïgnard* and *La Rôtisserie de la Reine Pédauque* (both 1893). The same attitudes pervaded the first three volumes of his *Histoire contemporaine,* a novel sequence impartially mocking church, right-wing opposition, and corrupt political establishment. But in the fourth volume, *Monsieur Bergeret à Paris* (1901; *Monsieur Bergeret in Paris*), France's commitment to the pro-Dreyfus faction in the Dreyfus Affair introduced both a more bitter note to his satire and an unaccustomed commitment to positive humanitarian ideals. Like many other Dreyfusards he was to be disillusioned by the aftermath of the affair, a response typified by his extended satire of French society through the ages in *L'Île des Pingouins* (1908; *Penguin Islands*) and his condemnation of fanaticism in his novel on the French Revolution, *Les Dieux ont soif* (1912; *The Gods Are Athirst*). (C.Ro.)

Beginning the 20th century: from 1900 to 1940

THE LEGACY OF THE 19TH CENTURY

French writing of the first quarter of the 20th century reveals a dissatisfaction with the pessimism, skepticism, and narrow rationalism of the preceding age and displays a new confidence in man's possibilities, although this is undercut by World War I. There is a continuity with the poetry of the late 19th century but a rejection of the prose. Mallarmé and Rimbaud were models for Paul Valéry and Paul Claudel, but the Naturalist novel was considered by the new generation to be unduly deterministic and falsely objective. Charles-Louis Philippe, in *Bubu de Montparnasse* (1901; Eng. trans., *Bubu de Montparnasse*), used the material of Zola, namely the Paris slums, but depicted it through a narrator who could identify with the working-class protagonists.

In philosophy the rationalism of Taine and Renan exerted less influence than Henri Bergson's belief in man's creative ability and in intuition as an instrument to understand the universe. Among foreign thinkers Arthur Schopenhauer, so important to the preceding generation, gave way to Friedrich Nietzsche, whose books were read not for the superman theme but as a protest against the limitations of the mechanistic world.

Effect of the Dreyfus Affair

Literature continued to follow the political and social struggles of the Third Republic. The Dreyfus Affair unleashed a wave of "commitment," with Maurice Barrès and Charles Maurras on the anti-Dreyfus side (those opposed to a retrial for Dreyfus) and Charles Péguy on the other. The debate about the education system and about the separation of church and state form the backdrop to a Roman Catholic renaissance that begins with Claudel and Péguy and later finds expression in François Mauriac and Georges Bernanos. Although often politically conservative and theologically orthodox, this writing is audacious in depicting the individual's need for God and his personal quest for salvation. Meanwhile the anti-German sentiment, which stemmed from the 1870 defeat and was revived in the years immediately preceding World War I, helped create the Action Française, led by Maurras. The group sought to steer French culture toward integral nationalism, a policy that sought to restore the monarchy, even as the burgeoning Socialist movement also won adherents.

THE NOUVELLE REVUE FRANÇAISE AND ITS WRITERS

Although the Third Republic governments were weak, centrist coalitions that writers found difficult to admire, the hidden stability of French society made possible a literature that exalted individual experience. Some of the leading writers of the years before 1914 gathered around the *Nouvelle Revue Française,* which was founded by André Gide in 1908. The review, which became France's leading literary magazine while also spawning the Gallimard publishing house, shunned extremes, was distrustful of commitment, and sought a balance between modernity and tradition. Although its articles represented a network of dialogues rather than one fixed position, it nonetheless emphasized the authenticity of inner life.

Valery Larbaud's *A.O. Barnabooth: son journal intime* (1913; *A.O. Barnabooth: His Diary*) depicts the slow discovery of the self after an initial liberation. The enormously successful *Le Grand Meaulnes* (1913; *The Wanderer*) by Alain-Fournier (pseudonym of Henri-Alban Fournier) explored the new theme of adolescence; in poetry, Saint-John Perse (pseudonym of Alexis Léger) depicts the triumphant recovery of childhood in *Éloges* (1911; *Éloges, and Other Poems*); and Jacques Rivière's essays on painting, the Russian ballet, and contemporary writers show an excellent critical mind piecing together its age.

Novels of André Gide

The house of Gallimard published the four greatest writers of this period: Gide, Marcel Proust, Claudel, and Valéry. Gide's *Les Nourritures terrestres* (1897; *Fruits of the Earth*) and *L'Immoraliste* (1902; *The Immoralist*) encouraged a generation of French youth to question the values of family and tradition and to be guided by that part of themselves that typically was ignored or repressed by society. *Les Caves du Vatican* (1914; *The Vatican Swindle*) depicted the *acte gratuit,* which is undertaken not for gain or self-interest but because it expresses the deepest strain in one's character. Less influential was *La Porte étroite* (1909; *Strait Is the Gate*), which explains the concomitant themes of asceticism and sacrifice.

Pursuit of freedom also guided Gide's technical experiments from *Paludes* (1895; *Marshlands*), considered to be one of his most important books, to *Les Faux-Monnayeurs* (1926; *The Counerfeiters*). He rejected the closed world of the 19th-century novel and instead sought forms of narration that were both internal and multiple and that allowed the reader a greater range of interpretation. Although some authors continued to produce traditional novels and multivolume works (such as Romain Rolland's *Jean Christophe,* 1904–12) using a conventional plot and time sequence, the best writing of the period is consciously avant-garde.

Marcel Proust's *À la recherche du temps perdu* (1913–27; *Remembrance of Things Past*) subjects existence to a remorseless analysis that seems to be in the best tradition of the French psychological novel. In the best known episode, that of Swann's love affair with Odette, Proust shows how the object of love is an illusion created and defined by the lover. But if reason dissects, unconscious memory may create the totality of experience. In the last volumes of this work its structure, which is based on recurring motifs perceived by intuition, is laid bare.

Proust, Claudel, and Valéry

Unlike his great rival Gide, Paul Claudel believed that man's deepest self emerged in the dialogue with a God who is jealous and remote but never absent. Less pious and less of a preacher than he is sometimes considered to be, Claudel depicts in *L'Annonce faite à Marie* (1912; *Tidings Brought to Mary*) a heroine who restores the divine order by her suffering and isolation. Yet in *Partage de Midi* (1906; *Break of Noon*) sexual love, albeit purified and punished, is the image of God's love for man.

Claudel's masterpiece is *Le Soulier de satin* (1930; *The Satin Slipper*), in which divine grace haunts the characters who try in vain to escape it. This view of human character was well suited to the theatre, and Claudel's plays are full of buffoonery, ritual, and movement. They break with the realistic theatre of the 19th century, and they have also lasted better than the vaguely Symbolist plays of Maurice Maeterlinck, whose *Oiseau bleu* (1908; *The Blue Bird*) was more popular at the time.

Although he was also a great poet, Claudel may be less important than Paul Valéry, who inherited Mallarmé's sense that the language of poetry was separate from ordinary discourse. Valéry, however, pursued poetry not as an end in itself but as a form of self-knowledge. The act of writing and of reading poetry reveals various states of consciousness, and the process makes possible glimpses of totality. In his greatest poem, "Le Cimetière Marin" (1920; "The Graveyard by the Sea"), Valéry combines extraordinary self-awareness with sensuous enjoyment of the outside world. Earlier, in a prose work called "La Soirée avec Monsieur Teste" (1896; "An Evening with Monsieur Teste"), which offers parallels with *Paludes*, Valéry had parodied the ideal of complete self-awareness.

THE IMPACT OF WORLD WAR I

The confidence displayed in the pages of the *Nouvelle Revue Française* was sapped by the slaughter in the trenches in World War I. The influence of the Action Française was initially increased by the Allied victory, and antiwar novels, such as Henri Barbusse's *Le Feu* (1916; *Under Fire*), were relatively few in number. The early 1920s were a brilliant period, during which the cosmopolitanism of reviews like *Commerce*, directed by Valéry, Larbaud, and the poet Léon-Paul Fargue and including texts from many countries, was a conscious attempt to overcome the rifts created in Europe by the war. French writing was influenced by foreign artists like James Joyce, whose inner monologue was adapted into French by Larbaud in *Amants, heureux amants* (1923; "Lovers, Happy Lovers"); in turn, Paris drew writers like T.S. Eliot, Eugenio Montale, and William Carlos Williams. Cautiously Gide and Rivière began to defy the Action Française and to advocate more gentle treatment of the defeated Germany. A cult of travel writing was exemplified by Paul Morand, whose *Ouvert la nuit* (1922; *Open All Night*) and *Fermé la nuit* (1923; *Closed All Night*) were popular.

The Surrealists

A different note was struck by the Surrealists, who drew on and retrospectively exalted Guillaume Apollinaire. The poems of *Alcools* (1913; Eng. trans., *Alcools*) and Apollinaire's interest in Cubist painting were filtered through the imported slogans of Dadaism and turned into a movement that wished to break violently with existing French culture, including the masters of the *Nouvelle Revue Française*. By emphasizing the subconscious and by using devices such as automatic writing, the Surrealists sought to destroy both existing social structures and the ordered text. André Breton, Louis Aragon, and Paul Éluard led a movement that mocked the homeland, the army, the existing political parties, and ordinary methods of perception.

POLITICAL COMMITMENT

In their hatred of contemporary French society the Surrealists endorsed Communism, although only Louis Aragon played a significant role within the Communist Party. Political commitment became more important (even though Julien Benda had warned of its dangers in *La Trahison des clercs* [1927; *The Great Betrayal*]) in the early 1930s, both because Hitler's attainment of power in Germany increased the possibility of a Fascist Europe and because the stability of the Third Republic was undermined by economic depression. The political polarization that stemmed from events such as the Stavisky affair (1933–34), which led to charges of widespread corruption in the parliamentary regime, or the outbreak of the Spanish Civil War in 1936 caused writers to take political stances, whether on the right or the left. Meanwhile, discontent with both the paper revolutions of Surrealism and the individualism of the older *Nouvelle Revue Française* drove the next generation to discover itself through action.

On the political right, the best known protagonist was Pierre-Eugène Drieu la Rochelle, whose *Gilles* (1939) remains an important book because it depicts the evolution of a young man during the interwar years. The novel begins with Gilles's experience of the trenches and takes him through periods of despair until he decides that he must fight for General Franco in Spain. An avowed Fascist with anarchist tendencies, Drieu, who was to collaborate during the Nazi occupation of France, was followed by younger men such as Robert Brasillach, author of *Notre avant-guerre* (1941), and Lucien Rebatet, who, like Brasillach, contributed during the Occupation to the virulently anti-Semitic newspaper *Je Suis Partout*.

On the political left, Joseph Stalin's decision to end the policy of hostility toward the Socialist Party and instead to form Popular Fronts brought many writers into or close to the Communist Party. Newspapers like *Commune* and *Monde*, which advocated that literature should serve the cause of working-class liberation, were influential. Meanwhile Gide's adherence to and defection from Communism, depicted in *Retour de l'U.R.S.S.* (1936; *Return from the U.S.S.R.*), were widely discussed.

The books of Paul Nizan, who had joined the party earlier, exemplify this trend. *La Conspiration* (1938) is a satire on both the middle-class family and immature revolutionaries, while *Le Cheval de Troie* (1935; *Trojan Horse*) depicts a group of Communists engaged in the political struggle.

The issue of colonialism

An odd but important book is Henry de Montherlant's *La Rose de sable* (written in 1932 although not published until 1967). Montherlant offers a critique of French colonialism, using as his vehicle the figure of a nationalist officer who loses his belief in French rule over Morocco. The issue of colonialism had been posed by Gide in his *Voyage au Congo* (1927; *Travels in the Congo*), and it was to recur in the work of Albert Camus, who for a time was active in the Algiers cultural agency set up under the Popular Front by the newspaper *Commune*.

THE PAMPHLET

The bitter ideological clashes and the growing emotional tension brought a rush of political pamphlets. In France authors typically used the genre as a forum for personal opinion: an author would select a target and, in his pamphlet, pour out a diatribe of hatred against it. One example of "leftist" pamphleteering is Nizan's *Les Chiens de garde* (1932), in which the author excoriates Sorbonne philosophy, and another is Bernanos' *Les Grands Cimetières sous la lune* (1938; *A Diary of My Times*), which denounces Franco's Falangists. Authors on the political right were more prolific. An example is Henri Béraud's *Faut-il reduire l'Angleterre en esclavage?* (1935; "Must England Be Reduced to Slavery?"), which vituperates against perfidious Albion. But the dubious distinction of being the greatest pamphleteer falls to Louis-Ferdinand Céline (Louis-Ferdinand Destouches), whose *Bagatelles pour un massacre* (1937; "Trifles for a Massacre") and *L'Ecole des cadavres* (1938; "School for Corpses") are a frenzy of anti-Semitism and of pacifism at all costs.

THE NOVEL BETWEEN THE WARS

With a few notable exceptions, novels were less technically innovative in the second quarter of the century. The Surrealists made their contribution with André Breton's *Nadja* (1928; Eng. trans., *Nadja*) and Louis Aragon's *Le Paysan de Paris* (1926; *The Night-Walker*), and they influenced the works of Raymond Queneau. The tradition of the family novel was continued by Roger Martin du Gard's *Les Thibault* (1922–40), while Jules Romains delved into the history of the Third Republic for *Les Hommes de bonne volonté* (1932–47; *Men of Goodwill*). A popular adventure novel was Antoine de Saint-Exupéry's *Vol de nuit* (1931; *Night Flight*). More important are the novels of François Mauriac, whose *Noeud de vipères* (1932; *Vipers' Tangle*) and *Thérèse Desqueyroux* (1927; Eng. trans., *Thérèse Desqueyroux*) deploy the traditional form of the French psychological novel in order to depict characters who are deprived of God's grace and stranded in the desert of provincial middle-class society. Mauriac does not have the triumphant tone of earlier Roman Catholic writing and he analyzes the problems of guilt, spiritual loneliness, and confession.

The current of working-class writing, inaugurated by Philippe, was continued by Louis Guilloux. His *Maison du peuple* (1927) depicts a child's view of the oppressed but rebellious community in which he grows up, and his *Sang noir* (1935; *Bitter Victory*) is a bleaker depiction of provincial life. The regional novel, in fact, is well represented between the wars. Marcel Aymé gave it a satirical twist in *La Jument verte* (1933; *The Green Mare*), while Jean Giono advocated a return to the traditional values of peasant life in *Regain* (1930; *Harvest*). Realism also reasserted itself in the 1930s, and a popular success was Eugène Dabit's *Hôtel du nord* (1929; *Hotel du Nord*), which depicts the grim streets of northern Paris.

The novels of Malraux

Young French readers turned more avidly to the novels of André Malraux, whose novels depict the themes of adventure, Bolshevism, and revolutionary fraternity. *La Condition humaine* (1933; *Man's Fate*), which depicts the Communist uprising in Shanghai in 1927, shows a group of comrades who overcome their obsession with death by taking revolutionary action. *L'Espoir* (1937; *Man's Hope*) is a complex work that combines journalistic techniques, outbursts of lyricism, and switches in the narrative structure in order to depict the Spanish Civil War. If the book is in part an epic whose collective hero is the Republican fighters, it is also a debate about moral values versus efficiency and about the psychology of the Communist militant.

The best novelist of the 1930s, however, was Céline. *Voyage au bout de la nuit* (1932; *Journey to the End of the Night*) is a journey of discovery that moves from World War I, via Africa and America, to the drab suburbs of northern Paris. Intent on confronting the bleakest aspects of the human condition, Céline also wished to render them in a more immediate and brutal manner, so *Mort à credit* (1936; *Death on the Installment Plan*) is written in a tide of slang and obscenity that also possesses its own lyricism. By dismembering the sentences, scrapping plot, and moving without transition from reality to dream, Céline made the most original contribution of the interwar years to the art of novel writing.

POETRY

Valéry, Claudel, and Léon-Paul Fargue continued writing throughout this period and Saint-John Perse produced a great work, *Anabase* (1924; *Anabasis*). The influence of the Surrealists was strongest in poetry. Breton, whose work is rhetorical and only superficially chaotic, Aragon, and Paul Éluard (pseudonym of Eugène Grindel) all produced poetry that is still read. On the fringes of the movement were two better poets: Henri Michaux, whose prose poems *La Nuit remue* (1934; *The Night Moves*) are a striking example of that difficult genre, and René Char. Char's work exalts the mythical forces that reside in the countryside of southern France, with its bare hills and its twisted vegetation; of all Surrealists he has had the greatest influence on younger French poets. A simpler vein was tapped lyrically by Robert Desnos and satirically by Jacques Prévert, whose doggerel ballads reveal an anarchist's temperament.

THEATRE

Although the Surrealists retrospectively exalted Alfred Jarry's *Ubu roi* (1896; Eng. trans., *Ubu roi:* "King Ubu"), which exploded all theatrical conventions, they did not themselves produce plays of lasting value. In general the period between the wars was not a great era for the French theatre. The dominant playwrights were men like Édouard Bourdet, whose highly successful works were witty and clever but slight. A similar but better dramatist was Marcel Pagnol, whose *Marius* (1929; Eng. trans., *Marius*) is set in the Marseille docks. Cocteau and Jean Giraudoux were more serious figures, and Giraudoux's *La Guerre de Troie n'aura pas lieu* (1935; adapted in English as *Tiger at the Gates*) is characteristic of his fanciful manner.

The popular comedies of the day were anathematized by Antonin Artaud, but his theatre of cruelty was not to exert much influence until after World War II. Similarly the next generation would admire such directors as Jacques Copeau, whose Vieux-Colombier theatre had grown out of the *Nouvelle Revue Française;* Charles Dullin, who influenced the young Sartre; and Jean-Louis Barrault.

THE EVE OF WORLD WAR II

New influences were at work from the mid-1930s onward: the novels of the U.S. writers William Faulkner and John Dos Passos as well as the philosophy of Edmund Husserl and Martin Heidegger found an audience in France. Albert Camus published *L'Envers et l'endroit* (1937; *Betwixt and Between*) and *Noces* (1939; *Nuptials*), the two volumes of essays that reveal his sense of the beauty and the emptiness of life near the Mediterranean. Jean-Paul Sartre unravelled the French psychological novel and the traditional diary form in *La Nausée* (1938; *Nausea*), while in *Le Mur* (1939; *The Wall*) he depicted the aberrations that can occur in any person's character. His interest in Phenomenology, his sense of consciousness as a perpetual flight, and his insistence on the priority of existence over essence led him to set out the philosophy of Existentialism, which dominated the 1940s. (P.McC.)

Continuing the 20th century: from 1940

Writers continued to enjoy a leading place in cultural life, but the outstanding names were not mainly connected with imaginative literature. Their work was philosophical and critical and the movements with which they were associated extended beyond the field of literature as narrowly defined. Even within this field the *nouveau roman* (New Novel) was to challenge established conventions, and fiction no longer appeared to have a privileged place in literary culture. But, paradoxically, by asserting the primacy of the text as an object in itself, critics severed the last remaining links between its reality and the one it purported to describe, to conclude that all writing was essentially an exercise in the creation of a fictional "textual world."

THE GERMAN OCCUPATION AND POSTWAR FRANCE

France's defeat by German troops in 1940, and the resultant division of the country, was experienced as a national humiliation, and all Frenchmen and Frenchwomen were confronted with an unavoidable choice. Some writers escaped the country to spend the remaining years of the war in exile or with the Free French Forces. Others, because of political options made during the previous decade, moved directly into collaboration. And still others, because of pacifism or a belief that art could remain aloof from politics, tried to carry on as individuals and as writers, ignoring the taint of passive collaboration with the occupying forces or the Vichy government. Jean Cocteau and Jean Giono were among those and later were criticized for their conduct. Giono in fact was briefly imprisoned, as was Louis-Ferdinand Céline, whose reputation was seriously damaged by his anti-Semitism.

Several writers joined the military, as well as the intellectual, resistance. André Malraux served on many fronts and commanded a group of underground resistance fighters in World War II in France, confirming the image of the writer as a man of action; he was to serve as a minister under Charles de Gaulle in the postwar government and the Fifth Republic.

The German attack on the Soviet Union in 1941 was decisive for the French Communist Party, which was to gain considerably through its organized opposition to Fascism. The events of the 1930s and '40s strengthened the conviction that intellectuals could not remain politically uncommitted; the war clarified choices and made them seem crucial for the individual. After 1945, Existentialism, depicting mankind alone in a godless universe, rationalized this view of individuals as free to determine themselves through such choices.

Poetry of resistance

Meanwhile, the occupation brought prestige and an attentive audience to writers who upheld the honour of their defeated country. The poetry of resistance reached a wide public, notably in the works of Paul Éluard and Louis Aragon. Both were Communists (Aragon was to become the country's "cultural commissar" after 1945) and had been associated with the Surrealist movement from the 1920s. They turned in the war years to writing poems in direct language, and their poems were often transmitted orally through the occupied zone. A flourishing clandestine press was able to issue some publications, including the newspaper *Combat* and the Editions de Minuit, which brought out as its first book the story *Le Silence de la mer* (1942) by Vercors (Jean-Marcel Bruller). With the additional stories later added to a collection under the same title, this is probably the literary work that most accurately evokes the atmosphere and the dilemmas of the time. Camus's fable *La Peste* (1947; *The Plague*), an allegory set in a city visited by the plague, gave a universal dimension to the treatment of these dilemmas.

The war transformed the literary scene, eclipsing some writers and lending prestige to those who had made the fortunate political choices. During the occupation Sartre further elaborated his Existentialist philosophy in plays, such as *Les Mouches* (1943; "The Flies") and *Huis-Clos* (1944; *No Exit*, or *In Camera*) and expounded it in the treatise *L'Être et le néant* (1943; *Being and Nothingness*). After the liberation, the writer and his ideas set the tone for a postwar generation that congregated in the cafés and cellar clubs of Saint-Germain-des-Prés. The myth of this disillusioned youth, its district of Paris, its innocence, its jazz clubs, and its worship of Sartre were captured in Boris Vian's *L'Écume des jours* (1947).

Influence of Sartre

Sartre's reputation fluctuated widely during the 30 years from 1950 until his death: at times he was compared to Voltaire, and at other times he was dismissed as a senile fellow traveller. Until 1952 his name was linked with that of Camus, whose novel *L'Étranger* (1945; *The Stranger*, or *The Outsider*) expressed a vision similar to that of the early Sartre. But after their highly publicized break, Sartre moved toward the Existentialist Marxism of his *Critique de la raison dialectique* (1960; *Search for a Method*), and Camus toward a stoical humanism, his later fiction (*La Chute*, 1956; *The Fall*) showing evidence of his isolation, his creative unease, and his distress over France's war with Algeria.

The conflicts submerged in the euphoria of liberation surfaced during the Cold War period and were intensified by the colonial wars of the 1950s. Sartre's lifelong companion, Simone de Beauvoir, vividly depicted the contrary attractions of Communism and the United States for French intellectuals in her novel *Les Mandarins* (1954; *The Mandarins*). However, her analysis of the feminine condition, *Le Deuxième Sexe* (1949; *The Second Sex*), although reviled on its first appearance, was to be a more influential achievement, its themes being substantiated by her later autobiographical works. After Sartre's death she gave a moving account of his later years in *La Cérémonie des adieux* (1981; *Adieux, A Farewell to Sartre*), which was acknowledged as confirmation of their crucial role in intellectual life.

THE NOUVEAU ROMAN

The literary event of 1954 was *Bonjour tristesse*. Published when its author, Françoise Sagan (pseudonym of Françoise Quoirez), was only 19, this novel of adolescent love was written with "classical" restraint and a tone of cynical disillusionment and showed the persistence of traditional form in fiction. The previous generation, caught up in politics, had experimented with Socialist Realism (for example, the novels of Roger Vailland), yet, while the 1950s seem in retrospect to have been dominated by concern with form, the novel-reading public remained largely untouched by those experiments. The Naturalist novel survived in the work of Henri Troyat and others, while its assumptions about the role of the author and the nature of fictional "reality" continued to be taken for granted by a host of novelists and their readers.

On the other hand, these assumptions had long been challenged before the emergence of the *nouveau roman* (New Novel) in the work of Alain Robbe-Grillet, Claude Simon, Nathalie Sarraute, Michel Butor, and Marguerite Duras (pseudonym of Marguerite Donnadieu). What was new about these novelists, apart from the label applied to them, was their systematic rejection of the traditional framework of fiction—chronology, plot, character—and of the omniscient author. In place of these reassuring conventions, they offer texts that demand more of the reader, who is presented with compressed, repetitive, or only partially explained events from which to derive a meaning that will not, in any case, be definitive. In Robbe-Grillet's *La Jalousie* (1957; *Jealousy*), for example, the narrator's suspicions of his wife's infidelity are never confirmed or denied, but their obsessive quality is conveyed by the replacement of a chronological narrative with the insistent repetition of details or events. In *Le Libéra* (1968) by Robert Pinget there is no single narrator, while in the later novels of Jean Cayrol the narrative emanates from the sea, a field, the desert.

The *nouveau roman* was open to influence from works being written abroad (notably the work of William Faulkner) and from the cinema (Robbe-Grillet and Duras contributed to the *nouvelle vague*, or New Wave, style of filmmaking). But, by the time Robbe-Grillet's *Pour un nouveau roman* (*Toward a New Novel*) appeared in 1963, it was clear that the term covered a variety of approaches. In the same year, the Prix Théophraste Renaudot was awarded to Jean-Marie Le Clézio for *Le Procès-verbal* (*The Interrogation*), a novel welcomed partly because it was both "modern" and accessible. It was also heralded (prematurely) as offering an escape from the *nouveau roman*'s overintellectuality.

EXPERIMENTS IN THEATRE

During the 1940s the theatre provided a forum for the veiled expression of political dilemmas through the dramatization of individual crises of conscience. The plays of Montherlant, Jean Anouilh, Giraudoux, and Sartre made this, surprisingly, an outstanding moment for the Parisian stage. A government policy to provide state financial aid after the war led to the encouragement of great drama in the provinces (the Avignon Festival started in 1947) and the establishment of remarkable and innovative theatre companies, such as the Théâtre National Populaire and the Compagnie Jean-Louis Barrault–Madeleine Renaud. But, while the works they performed were inventive and intellectually stimulating, there was little change in the concept of theatre itself; this was a writer's theatre, the genre and the rules of stagecraft seemingly fixed within an agreed perception of what constituted "reality."

Jarry, Cocteau, and others had attempted to extend this perception long before in works that had opened drama to the irrational, the surreal, and the influence of circus and music hall. Artaud's *Le Théâtre et son double* (1938; *The Theatre and Its Double*) called for theatre that would shock its spectators into seeing the baseness of the real world. But it was only after the war, starting with plays staged in the Parisian fringe theatre, that the accepted notion of theatre and the primacy of the text were seriously challenged.

While the plays of Anouilh and Sartre conveyed their intentions effectively from the printed page, those of Jean Genet, Eugène Ionesco, Authur Adamov, and Samuel Beckett only revealed their full meaning in actual performance, the text often seeming flat, repetitive, or banal until brought to life on the stage. Though Genet's *Les Bonnes* (*The Maids*) appeared in 1947 and Ionesco's *La Cantatrice chauve* (*The Bald Soprano*) in 1949, public recognition of the new theatre did not come until 1953, with Roger Blin's production of Beckett's *En attendant Godot* (1952; *Waiting for Godot*). Originating in the fool of Shakespearean drama and the tramp of silent comedy, the characters of Vladimir and Estragon convey the absurdity and the tragedy of existence in an artistic context that resembles musical composition more than the classical "imitation of nature."

POSTWAR POETRY

Developments in the novel and the theatre were easier to define than those in poetry, where the lack of a broad readership was, in itself, an encouragement to fragmentation. The works of Jacques Prévert and the songs of Georges Brassens and Jacques Brel did achieve the status of popular poetry; but, apart from Saint-John Perse, there was no major figure in the tradition of Claudel and Valéry, and the poetry of the post-Surrealist generation appeared to have no clear formal or ideological direction. The mainstream of French poetry was represented by René Char, Pierre Emmanuel (pseudonym of Noël Mathieu), and Yves Bonnefoy. In contrast to the tendency to abstract and symbolic language that characterized their poetry, however, Francis Ponge in *Le Parti pris des choses* (1942) and later works used wordplay and devices to emphasize the act of writing, in prose poems centred on the impersonal description of objects.

On the whole, the intellectual bourgeoisie that might have provided the audience for poetry looked to the written word for the expression of ideas and found aesthetic stimulation in the visual arts, especially the cinema. A younger generation, from the late 1960s, was more open to fantasy and the imagination, but impatient of formal discipline. The "do-it-yourself" poetry that appealed to this group's egalitarian instincts was as ephemeral as the little magazines in which it appeared during the 1970s, and the "crisis of verse" that Jacques Roubaud described in *La Vieillesse d'Alexandre* (1978) remained unresolved.

Roubaud's own poetry, including *Trente et un au cube* (1973), looked to Japanese literature as the inspiration for work that was structured, yet free from the burden of European rhetoric. He was associated with the writers of OuLiPo (Ouvroir de Littérature Potentielle) who, inspired by Alfred Jarry and by Raymond Queneau, sought during the 1970s to escape from the vague "poeticism" of much contemporary free verse through the acceptance of rigorous formal constraints. Their fondness for wordplay and sometimes unbelievably demanding forms was illustrated in the works of Georges Perec. As well as poetry, Perec wrote novels, including *La Disparition* (1969), a text composed entirely without using the letter *e*, and *La Vie mode d'emploi* (1978), his most accessible work, built around the inhabitants of an apartment house.

Playful aspects of literary composition

This renewal of interest in the playful aspects of literary composition was consistent with contemporary critical theory and was accompanied by a reassessment of some earlier literature, such as the poetry of the *grands rhétoriqueurs*. Queneau, most widely known as the author of *Zazie dans le métro* (1959; *Zazie*), was the most immediate predecessor with his stylistic demonstrations in *Exercices de style* (1947) and the 10 sonnets of *Cent mille milliards de poèmes* (1961), which the reader was invited to rearrange in the hundred-thousand-billion ways indicated by the title. This unsolemn reaction to formlessness was clearly healthy, and the success of series devoted to the work of the more established poets showed that there was a readership even for quite demanding work, though that readership's social composition was less precisely defined and less coherent than it had been in earlier periods.

THE 1960S

In the early 1960s, free of colonial entanglements, France enjoyed increasing stability and affluence while, despite *le fast-food le marketing,* and *le rock,* French culture preserved an individual character and was defended against such transatlantic imports by René Etiemble in his polemic *Parlez-vous franglais?* (1964). The technocratic middle class, which benefitted most from the country's prosperity, was open to new ideas in science, and its materialist outlook found expression in *Le Hasard et la nécessité* (1970; *Chance and Necessity*) by Jacques Monod.

Monod, who won the Nobel Prize for Physiology or Medicine for 1965, rejected earlier ideologies, including religion, and drew on science for a view of the human place in the universe. The new technology seemed to promise endless growth and the erosion of class divisions. Perec, in *Les Choses* (1965), warned of the emptiness of this consumer society, while a bitter novel by Christiane Rochefort, *Les Petits Enfants du siècle* (1961), satirized the welfare state and revealed the drabness of life in a working-class housing complex. *La Dentellière* (1974), by Pascal Lainé, showed that the technocratic society had not abolished class differences and announced the next decade's concern for the position of women.

The most significant developments seemed to be outside the field of imaginative literature: in the structural anthropology of Claude Lévi-Strauss, the semiology of Roland Barthes, the neo-Freudian psychology of Jacques Lacan, and the philosophy of Michel Foucault or Jacques Derrida. Despite *Time* magazine's disparagement in 1969 of French culture, it can be argued that French writers, more than those of any other country, were making original contributions to almost every field of social science and the humanities.

Structuralism

Their work transcended academic boundaries. Structuralism, reacting against the Phenomenology of Sartre, asserted the crucial importance of relationships between phenomena, while semiology analyzed systems of signs, notably language, through which one attributes meaning to the raw data of reality. The journal *Tel Quel*, edited by Philippe Sollers, was the focus for a theory of writing as a concept to be extended beyond the realm of literature.

The New Criticism demanded more of the reader, who was to become an active participant in decoding the text, not a passive recipient. Critics emphasized context and variant readings, none of which was to be exclusively "correct." The text itself became the object of closer and closer scrutiny: a single line from a poem by Laforgue could give rise to five pages of exegesis in a critical journal.

The New Critics despised the university establishment and met with opposition from it about the time of Barthes's *Sur Racine* (1963; "On Racine"). The educational system was rigid and outdated. A liberal university admissions policy was combined with a teaching method based largely on formal lectures, and the vast student body found itself with no say in the running of a system that seemed to be largely irrelevant to its needs.

THE EVENTS OF 1968

During the student revolt in May 1968, streets, factories, schools, and universities became the stage for a spontaneous performance, the decor provided by posters and graffiti elevated to a popular art form. The theatre itself in the late 1960s and the 1970s moved still further away from the "writer's theatre" of earlier times, experimenting with audience participation and improvisation. Rock music and comic books flourished, and television, closely controlled by the government under de Gaulle, began to play an increasing role in cultural life; discussion programs and spin-offs from serials or adaptations were taking over from the press in guiding taste.

The events of 1968 led to no conclusions, and the community's reaction emphasized continuity and conservatism. The academies, the literary prizes, the dominant publishing houses, and the university elite survived more or less intact. The most evident long-term benefit of the upheaval was the encouragement it gave to the feminist movement. The analysis begun by Simone de Beauvoir was taken up in a host of works, and the novels of Claire Etcherelli, Marie Cardinal, Chantal Chawaf, and Hélène Cixous were among many that, in different ways, affirmed the value of feminine experience. Historians of both sexes were stimulated to reassess past roles of women, the family, and sexuality. The gains, by the 1980s, were enormous, though Marguerite Duras, whose novels had long emphasized the freeing of their women characters from a life-destroying social and domestic background, dissociated herself from some aspirations of the feminist movement, which she also found restricting.

LITERATURE AFTER 1970

With no movement obviously set to replace the *nouveau roman* in fiction, there were dire predictions about the future of literary culture. Only Michel Tournier was generally agreed to be outstanding. *Le Roi des Aulnes* (1971) was an extraordinary combination of myth and parable, and Tournier's imagination, as well as his skillful narrative technique, were confirmed in *Les Météores* (1975) and *Gilles et Jeanne* (1983). But critical acclaim for Tournier's work did not encourage imitation, and the most one could say of the novel up to the mid-1980s was that it exhibited a variety of tendencies, from post-*nouveau roman* experiments to traditional survivals.

The frustrations of the times may have added to the attraction of the historical novel. Marguerite Yourcenar (pseudonym of Marguerite de Crayencour), who in 1980 became the first woman elected to the Académie Française, had shown that the genre could appeal to more than escapism. *Mémoires d'Hadrien* (1951; *Memoirs of Hadrian*) and *L'Oeuvre au noir* (1968; *The Abyss*) were portraits of men who overcame the limitations of their time, as well as rich evocations of the past. History proved able to accommodate a vast range of fiction, from popular romance (Jeanne Bourin's best-selling *La Chambre des dames,* 1979) and fictionalized biography to the linguistic and narrative experiments of Claude Simon (*Les Géorgiques,* 1981) and Pierre Guyotat (*Le Livre,* 1984) or the sensitive prose of Florence Delay (*L'Insuccès de la fête,* 1980) and Christiane Singer (*La Mort viennoise,* 1978). It was sustained by the prestige of historiography: Michel Foucault's studies of sexuality and attitudes toward death, and the social history associated with the journal *Annales,* edited by Emmanuel Le Roy Ladurie, whose *Montaillou* (1975) was an international best-seller.

There was a corresponding interest in biography, autobiography, and memoirs. The novelists Julien Green and Julien Gracq (pseudonym of Louis Poirier) were among several figures of an earlier generation who began to publish journals and memoirs rather than fiction. The public delighted in the stories of Marcel Pagnol's Provençal childhood, but also in Nathalie Sarraute's *Enfance* (1983; *Childhood*), and seemed increasingly prepared to come to terms with some technical innovations. Georges Perec's *W: ou, le souvenir d'enfance* (1975; *W; or, The Memory of Childhood*) is an autobiography formed of the alternating chapters of two seemingly unconnected texts, which eventually find their resolution in the concentration camp.

Detective fiction, a genre sometimes exploited by the *nouveau roman,* had an outstanding practitioner in Georges Simenon, who, during the 1970s, also turned to autobiography. The gangster novels of Albert Simonin made imaginative use of Parisian slang, but the chief attraction of the thriller for more "literary" writers was its form, which they, like a number of filmmakers, adopted as a framework for the investigation of questions of identity or moral and political dilemmas (for example, in Michel del Castillo's *La Nuit du décret,* 1981).

The period after 1968 was one of adjustment to political, economic, and social changes, exemplified by the development of the feminist movement. The analysis begun by Simone de Beauvoir in *The Second Sex* was taken up in the novels of Claire Etcherelli, Marie Cardinal, Chantal Chawaf, and Hélène Cixous. Cixous developed the idea of *écriture féminine*, which emphasized such "feminine" intellectual traits as openness and play, and expressed this theory in various prose fictions. Monique Wittig sought to model women's struggle for self-designation in such novels as *Le Corps lesbien* (1973; *The Lesbian Body*). In the theatre Marguerite Duras's *India Song* (1972; Eng. trans. *India Song*) found new configurations to express the protean nature of desire.

Approaching the 21st century: the 1980s and beyond

POSTCOLONIAL LITERATURE

Contributions by immigrants and their descendants

Pierre Nora, writing the closing essay to his great project of national cultural commemoration, *Les Lieux de mémoire* (1984–92; *Realms of Memory*), commented on the multiplicity of meanings and connotations that since the mid-1970s had become invested in the once-monolithic concepts of identity, memory, and patrimony: "the three key words of contemporary consciousness." An important contribution was being made to French cultural life not only by Francophone writers from North Africa, sub-Saharan Africa, and the Caribbean but by descendants of immigrants in France itself. Works produced by the children of North African immigrants began to be published in the early 1980s; their insights into cross-cultural identity and the patterns of life in underprivileged working-class suburbs enriched the literature in works such as Leïla Houari's *Zeida de nulle part* (1985; "Zeida from Nowhere") or her *Poème-fleuve pour noyer le temps présent* (1995; "Stream-of-Consciousness Poetry to Drown the Present In"). The French also began to come to terms with the Algerian conflict, as evidenced by the success in France of Albert Camus's posthumously published *Le Premier Homme* (1994; *The First Man*), an autobiographical novel based on his father's childhood in Algeria in a colonist milieu. Having established—in novels such as *L'Amour, la fantasia* (1985; *Fantasia: An Algerian Cavalcade*)—her reputation as both ardent defender and critic of her native Algeria, the acclaimed novelist Assia Djebar began to produce fictions that look to Algeria but are also alert to the hierarchies of power in Europe, as in *Les Nuits de Strasbourg* (1997; "Strasbourg Nights").

POSTMODERNISM AND MAINSTREAM PROSE FICTION

Thought and sensibility at the end of the century were governed by postmodernism. Jean-François Lyotard's *La Condition postmoderne* (1979; *The Postmodern Condition*) declared the end of the modes and concepts that had fueled 18th-century scientific rationalism and the industrial and capitalist society to which it gave birth: the "grand narratives" of historical progress and concepts of universal moral value and absolute worth. Societies were to be seen instead as collections of games or performances, played within arbitrary sets of rules. History, it seemed, had no more use, and value judgments were at an end.

Mingling art forms

The constructs of postmodernism, which encouraged the mingling of different art forms and genres, heavily affected the novel of this period. Jean Echenoz's comic pastiches of adventure, detective, and spy stories pleased both critics and the reading public. Photography and writing joined to produce the *photo-roman*, concerned with exploring the relationship between the image and the narrative work that goes into its construction and interpretation. Good examples of the *photo-roman* are Roland Barthes's *La Chambre Claire* (1980; *Camera Lucida*) and Hervé Guibert's *Vice* (1991). Gay writing found an important collective focus in the AIDS crisis, most notably in Guibert's best-selling *À l'ami qui ne m'a pas sauvé la vie* (1990; *To the Friend Who Did Not Save My Life*). Social issues were addressed in the autobiographical fiction of Annie Ernaux, who, in *La Place* (1983; *Positions*; also published as *A Man's Place*) and *Une femme* (1988; "A Woman"; Eng. trans. *A Woman's Story*), looked at the stresses between generations created by social change and changes of class allegiance. Christiane Rochefort's novel of child abuse, *La Porte au fond* ("The Door at the Back of the Room"), appeared in 1988. Hélène Cixous's feminist classic, *Le Livre de Prométhéa* (1983; *The Book of Promethea*), learned, funny, sparkling, and innovative, achieved its writer's ambition to make a distinctive model of the desiring feminine subject. Marguerite Duras's autobiographical novel *L'Amant de la Chine du Nord* (1991; *The North China Lover*) voiced its author's own version of the feminine erotic. Monique Wittig stylized lesbian sadomasochism in her parodic *Virgile, non* (1985; "Virgil, No"; Eng. trans. *Across the Acheron*). Another generation began publishing in the 1980s. Marie Redonnet's prose fiction sits at the edge of popular culture, in a bizarre blend of realism and fantasy. Chantal Chawaf's sensually charged prose offers a highly original version of the blood rhythms of the body in *Rédemption* (1989; Eng. trans. *Redemption*), a new kind of vampire novel.

Radically diverse accounts of life in the contemporary world emerged. Sylvie Germain's magic realism works on landscapes steeped in history. Her novel *La Pleurante des rues de Prague* (1992; *The Weeping Woman on the Streets of Prague*) is a dreamlike, surreal evocation of a city haunted by its sorrowful history. Her *Tobie des marais* (1998; *The Book of Tobias*) reworks the Apocryphal tale in a France that is simultaneously, and pleasingly, both medieval and modern. The narrative personae of Michel Houellebecq's highly successful novels *Extension du domaine de la lutte* (1994; *Whatever*) and *Les Particules élémentaires* (1998; *The Elementary Particles,* also published as *Atomised*) are computer-age Baudelaires, narcissistic and world-weary. Marie Darrieussecq's *Truismes* (1996; *Pig Tales: A Novel of Lust and Transformation*) is an imaginative political and moral satire depicting the blackly comic world of a young working woman with a highly materialistic lifestyle who begins to turn into a pig.

POETRY

Christian Prigent asked in an essay of 1996 what poets were good for in the modern world. His work and that of such well-established figures as Philippe Jaccottet (*La Seconde Semaison* [1996; "The Second Sowing"]) still reached audiences at the turn of the century, and Michel Houellebecq published his collected poems (*Poésies*) in 2000. Martin Sorrell's bilingual anthology, *Elles* (1995; "They [the women]"), showed the flourishing state of women's poetry. In it, poets such as Marie-Claire Bancquart, Andrée Chedid, and Jeanne Hyvrard offer their own insights into the problematic of gender roles and the challenge of finding a female poetic voice.

DRAMA

Theatre scripts revived

Most interesting of all, perhaps, was the revival of scripted drama at the end of the 20th century. The directors' theatre that held sway in the 1970s and early 1980s (inspiring spectacular and innovative staging developments and giving great scope to actors) had marginalized new writing. Ministry of Culture subsidies supported the work of Michel Vinaver and Bernard-Marie Koltès, whose plays are concerned with individuals struggling with the institutional discourses—family, law, politics—of which contemporary consumer society and their own identities are woven. The quick exchanges of Vinaver's play *L'Émission de télévision* (1990; "The Television Program") express the anxieties of a world in which realities are constantly shifting. Koltès's work is especially concerned with the fast-expanding numbers of the marginalized in a postcolonial world. His *Dans la solitude des champs de coton* (1986; "In the Solitude of Cotton Fields"), written two years before his death from AIDS and now translated and performed across the world, brilliantly embodies his central belief that modern life is focused on the deal; that is, the struggle for power between unequal individuals, client and dealer, seller and buyer. What is acted out on the Koltèsian stage are the rhetorical performances by which people live—on the edge of darkness, at the frontiers of disorder.

It is perhaps in the theatre that the value of current insights into the ludic and performative nature of the human condition can most easily be tested. At the close of the century, the most modern of creative writers in this respect remained Irish-born Samuel Beckett, who died in 1989 but whose importance, influence, and presence continued to grow. Exploiting and offsetting the rhythms of language, vision, and movement in order to explore the limits and the potential of form, Beckett's drama enshrines the serious nature of play. In so doing, it brings into focus what have always been the best parts of the French contribution to the Western cultural tradition: the analytical vision that penetrates the patterns and structures of the historical moment, the synthetic imagination that clarifies those patterns for others to see, in all their force and intensity—and the driving desire to see them otherwise. (R.C.Bu./J.Bir.)

BIBLIOGRAPHY. General literary histories include JENNIFER BIRKETT and JAMES KEARNS, *A Guide to French Literature: From Early Modern to Postmodern* (1997); PETER FRANCE (ed.), *The New Oxford Companion to Literature in French* (1995); DAVID HOLLIER (ed.), *A New History of French Literature* (1989, reissued 1994); and ANTHONY LEVI, *Guide to French Literature*, 2 vol. (1992–94).

Middle Ages: Histories include JOHN FOX, *The Middle Ages* (1974), vol. 1 in the series *A Literary History of France*, ed. by P.E. CHARVET. Among the better anthologies are C.W. ASPLAND (ed.), *A Medieval French Reader* (1979); and BRIAN WOLEDGE (ed.), *The Penguin Book of French Verse*, vol. 1 (1961). Studies on the epics and the romances are JESSIE CROSLAND, *The Old French Epic* (1951, reprinted 1971); PIERRE LE GENTIL, *The Chanson de Roland* (1969; originally published in French, 1955); JOSEPH J. DUGGAN, *A Guide to Studies on the* Chanson de Roland (1976); ROGER SHERMAN LOOMIS (ed.), *Arthurian Literature in the Middle Ages* (1959, reprinted 1979); L.T. TOPSFIELD, *Chrétien de Troyes: A Study of the Arthurian Romances* (1981); DOUGLAS KELLY, *Chrétien de Troyes: An Analytic Bibliography* (1976); and DAVID J. SHIRT, *The Old French Tristan Poems* (1980). Useful works on the lyric include L.T. TOPSFIELD, *Troubadours and Love* (1975, reprinted 1978); and FREDERICK GOLDIN (comp.), *Lyrics of the Troubadours and Trouvères* (1973, reprinted 1983). Among the volumes on prose literature are JANET M. FERRIER, *French Prose Writers of the Fourteenth and Fifteenth Centuries* (1966). GRACE FRANK, *The Medieval French Drama* (1954, reprinted 1972), is helpful.

The 16th century: A brief overview is given in A.J. KRAILSHEIMER (ed.), *The Continental Renaissance* (1971, reissued 1978). Among studies of the Pléiade is GRAHAME CASTOR, *Pléiade Poetics: A Study in Sixteenth-Century Thought and Terminology* (1964). Other works on the poetry of the time are TERENCE C. CAVE, *The Cornucopian Text: Problems of Writing in the French Renaissance* (1979, reissued 1985), and *Devotional Poetry in France c. 1570–1613* (1969).

The 17th century: Important works include P.J. YARROW, "The Seventeenth Century, 1715–1789," in P.E. CHARVET (ed.), *A Literary History of France*, vol. 2 (1967); JOHN CRUICKSHANK (ed.), *French Literature and Its Background*, vol. 2: *The Seventeenth Century* (1969); JOHN LOUGH, *An Introduction to Seventeenth-Century France* (1954, reissued 1969); W.D. HOWARTH, *The Seventeenth Century* (1965), vol. 1 in *Life and Letters of France*; and A.J. KRAILSHEIMER (ed.), *Studies in Self-Interest: From Descartes to La Bruyère* (1962). Among works on the drama are C.J. GOSSIP, *Introduction to French Classical Tragedy* (1981); GORDON POCOCK, *Corneille and Racine: Problems of Tragic Form* (1973); and H.T. BARNWELL, *The Tragic Drama of Corneille and Racine: An Old Parallel Revisited* (1982).

The 18th century: A standard work on the period is provided by ROBERT NIKLAUS, "The Eighteenth Century, 1715–1789," in P.E. CHARVET (ed.), *A Literary History of France*, vol. 4 (1967). Works on the novel include PETER BROOKS, *The Novel of Worldliness: Crébillon, Marivaux, Laclos, Stendhal* (1969); JOAN HINDE STEWART, *Gynographs: French Novels by Women of the Late Eighteenth Century* (1993); and VIVIENNE MYLNE, *The Eighteenth-Century French Novel: Techniques of Illusion*, 2nd ed. (1981).

The 19th century: Excellent studies for 1800 to 1850 appear in D.G. CHARLTON (ed.), *The French Romantics*, 2 vol. (1984); and W.D. HOWARTH, *Sublime and Grotesque: A Study of French Romantic Drama* (1975). Among the better studies of the literature from 1850 to 1900 are CHRISTOPHER ROBINSON, *French Literature in the Nineteenth Century* (1978); F.W.J. HEMMINGS, *Culture and Society in France, 1848–1898: Dissidents and Philistines* (1971); G.M. CARSANIGA and F.W.J. HEMMINGS (ed.), *The Age of Realism* (1974, reissued 1978); A.G. LEHMANN, *The Symbolist Aesthetic in France, 1885–1895*, 2nd ed. (1968, reprinted 1977); ANNA BALAKIAN, *The Symbolist Movement: A Critical Appraisal* (1967, reissued 1977); MARVIN CARLSON, *The French Stage in the Nineteenth Century* (1972); JENNIFER BIRKETT, *The Sins of the Fathers. Decadence in France 1870–1914* (1986); and RICHARD GRIFFITHS, *The Reactionary Revolution: The Catholic Revival in French Literature, 1870–1914* (1966).

The 20th century: An overview of the period to 1970 exists in the series *Littérature français* in PIERRE-OLIVIER WALZER, *Le XX*^{e} *siècle*, vol. 1: 1896–1920 (1975), and GERMAINE BRÉE, vol. 2: 1920–1970 (1978). The second volume has been translated into English: GERMAINE BRÉE, *Twentieth-Century French Literature*, trans. by LOUISE GUINEY, (1983). Works on the novel include HENRI PEYRE, *The Contemporary French Novel* (1955, reissued 1959); JOHN STURROCK, *The French New Novel: Claude Simon, Michel Butor, Alain Robbe-Grillet* (1969); CELIA BRITTON, *The Nouveau Roman: Fiction, Theory, Politics* (1992); EDMUND J. SMYTH (ed.), *Postmodernism and Contemporary Fiction* (1991); MARGARET ATACK and PHIL POWRIE (eds.), *Contemporary French Fiction by Women: Feminist Perspectives* (1990); EVA MARTIN SARTORI and DOROTHY WYNNE ZIMMERMAN (eds.), *French Women Writers. A Bio-Bibliographical Source Book* (1991). Among works on the theatre are MARTIN ESSLIN, *The Theatre of the Absurd*, 3rd ed., rev. and enlarged (1980, reissued 2001); and DAVID BRADBY, *Modern French Drama, 1940–1990*, 2nd ed. (1991). Poetry studies include MARCEL RAYMOND, *From Baudelaire to Surrealism* (1970; originally published in French, 1933); PETER BROOME and GRAHAM CHESTERS, *An Anthology of Modern French Poetry, 1850–1950* (1976); MICHAEL BISHOP, *The Contemporary Poetry of France: Eight Studies* (1985); MARTIN SORRELL (ed.), *Modern French Poetry: A Bilingual Anthology Covering Seventy Years* (1992) and *Elles: A Bilingual Anthology of Modern French Poetry by Women* (1995). CHRISTOPHER ROBINSON, *Scandal in the Ink: Male and Female Homosexuality in Twentieth-Century French Literature* (1995), is an excellent study.

(D.D.R.O./D.Mé./W.D.H./H.T.M./C.Sm./
C.Ro./P.McC./R.C.Bu./J.Bir.)

Freud

Sigmund Freud may justly be called the most influential intellectual legislator of his age. He was the founder of psychoanalysis, at once a theory of the human psyche, a therapy for the relief of its ills, and an optic for the interpretation of culture and society. Despite repeated criticisms, attempted refutations, and qualifications of Freud's work, its spell remained powerful well after his death and in fields far removed from psychology as it is narrowly defined. If, as the American sociologist Philip Rieff once contended, "psychological man" replaced such earlier notions as political, religious, or economic man as the 20th century's dominant self-image, it is in no small measure due to the power of Freud's vision and the seeming inexhaustibility of the intellectual legacy he left behind.

Mary Evans/Sigmund Freud Copyrights (courtesy of W.E. Freud)

Freud, 1921.

Early life and training. Freud was born on May 6, 1856, in Freiberg, Moravia (now Příbor, Czech.), then part of the Habsburg empire. His father, Jakob Freud, was a Jewish wool merchant who had been married once before he wed Freud's mother, Amalie Nathansohn. The father, 40 years old at Freud's birth, seems to have been a relatively remote and authoritarian figure, while his mother appears to have been more nurturant and emotionally available. Although Freud had two older half-brothers, his strongest if also most ambivalent attachment seems to have been to a nephew, John, one year his senior, who provided the model of intimate friend and hated rival that Freud reproduced often at later stages of his life.

The shaping influence of Vienna

In 1859 the Freud family was compelled for economic reasons to move to Leipzig and then a year after to Vienna, where Freud remained until the Nazi annexation of Austria 78 years later. Despite Freud's dislike of the imperial city, in part because of its citizens' frequent anti-Semitism, psychoanalysis reflected in significant ways the cultural and political context out of which it emerged. For example, Freud's sensitivity to the vulnerability of paternal authority within the psyche may well have been stimulated by the decline in power suffered by his father's generation, often liberal rationalists, in the Habsburg empire. So too his interest in the theme of the seduction of daughters was rooted in complicated ways in the context of Viennese attitudes toward female sexuality.

In 1873 Freud was graduated from the Sperl Gymnasium and, apparently inspired by a public reading of an essay by Goethe on nature, turned to medicine as a career. At the University of Vienna he worked with one of the leading physiologists of his day, Ernst von Brücke, an exponent of the materialist, antivitalist science of Hermann von Helmholtz. In 1882 he entered the General Hospital in Vienna as a clinical assistant to train with the psychiatrist Theodor Meynert and the professor of internal medicine Hermann Nothnagel. In 1885 Freud was appointed lecturer in neuropathology, having concluded important research on the brain's medulla. At this time he also developed an interest in the pharmaceutical benefits of cocaine, which he pursued for several years. Although some beneficial results were found in eye surgery, which have been credited to Freud's friend Carl Koller, the general outcome was disastrous. Not only did Freud's advocacy lead to a mortal addiction in another close friend, Ernst Fleischl von Marxow, but it also tarnished his medical reputation for a time. Whether or not one interprets this episode in terms that call into question Freud's prudence as a scientist, it was of a piece with his lifelong willingness to attempt bold solutions to relieve human suffering.

Freud's scientific training remained of cardinal importance in his work, or at least in his own conception of it. In such writings as his "Entwurf einer Psychologie" (written 1895, published 1950; "Project for a Scientific Psychology") he affirmed his intention to find a physiological and materialist basis for his theories of the psyche. Here a mechanistic neurophysiological model vied with a more organismic, phylogenetic one in ways that demonstrate Freud's complicated debt to the science of his day.

Jean-Martin Charcot

In late 1885 Freud left Vienna to continue his studies of neuropathology at the Salpêtrière clinic in Paris, where he worked under the guidance of Jean-Martin Charcot. His 19 weeks in the French capital proved a turning point in his career, for Charcot's work with patients classified as "hysterics" introduced Freud to the possibility that psychological disorders might have their source in the mind rather than the brain. Charcot's demonstration of a link between hysterical symptoms, such as paralysis of a limb, and hypnotic suggestion implied the power of mental states rather than nerves in the etiology of disease. Although Freud was soon to abandon his faith in hypnosis, he returned to Vienna in February 1886 with the seed of his revolutionary psychological method implanted.

Several months after his return Freud married Martha Bernays, the daughter of a prominent Jewish family whose ancestors included a chief rabbi of Hamburg and Heinrich Heine. She was to bear six children, one of whom, Anna Freud, was to become a distinguished psychoanalyst in her own right. Although the glowing picture of their marriage painted by Ernest Jones in his biography of Freud has been nuanced by later scholars, it is clear that Martha Bernays Freud was a deeply sustaining presence during her husband's tumultuous career.

Friendship with Wilhelm Fliess

Shortly after his marriage Freud began his closest friendship, with the Berlin physician Wilhelm Fliess, whose role in the development of psychoanalysis has occasioned widespread debate. Throughout the 15 years of their intimacy Fliess provided Freud an invaluable interlocutor for his most daring ideas. Freud's belief in human bisexuality, his idea of erotogenic zones on the body, and perhaps even his imputation of sexuality to infants may well have been stimulated by their friendship.

A somewhat less controversial influence arose from the partnership Freud began with the physician Josef Breuer after his return from Paris. Freud turned to a clinical practice in neuropsychology, and the office he established at Berggasse 19 was to remain his consulting room for almost half a century. Before their collaboration began, during the early 1880s, Breuer had treated a patient named Bertha Pappenheim—or "Anna O.," as she became known in the literature—who was suffering from a variety of hysterical symptoms. Rather than using hypnotic suggestion, as had Charcot, Breuer allowed her to lapse into a state resembling autohypnosis, in which she would talk about the initial manifestations of her symptoms. To Breuer's surprise, the very act of verbalization seemed to provide some relief from their hold over her (although later scholarship has cast doubt on its permanence). "The talking cure" or "chimney sweeping," as Breuer and Anna O., respectively, called it, seemed to act cathartically to produce an abreaction, or discharge, of the pent-up emotional blockage at the root of the pathological behaviour.

Technique of free association

Psychoanalytic theory. Freud, still beholden to Charcot's hypnotic method, did not grasp the full implications of Breuer's experience until a decade later, when he developed the technique of free association. In part an extrapolation of the automatic writing promoted by the German Jewish writer Ludwig Börne a century before, in part a result of his own clinical experience with other hysterics, this revolutionary method was announced in the work Freud published jointly with Breuer in 1895, *Studien über Hysterie* (*Studies in Hysteria*). By encouraging the patient to express any random thoughts that came associatively to mind, the technique aimed at uncovering hitherto unarticulated material from the realm of the psyche that Freud, following a long tradition, called the unconscious. Because of its incompatibility with conscious thoughts or conflicts with other unconscious ones, this material was normally hidden, forgotten, or unavailable to conscious reflection. Difficulty in freely associating—sudden silences, stuttering, or the like—suggested to Freud the importance of the material struggling to be expressed, as well as the power of what he called the patient's defenses against that expression. Such blockages Freud dubbed resistance, which had to be broken down in order to reveal hidden conflicts. Unlike Charcot and Breuer, Freud came to the conclusion, based on his clinical experience with female hysterics, that the most insistent source of resisted material was sexual in nature. And even more momentously, he linked the etiology of neurotic symptoms to the same struggle between a sexual feeling or urge and the psychic defenses against it. Being able to bring that conflict to consciousness through free association and then probing its implications was thus a crucial step, he reasoned, on the road to relieving the symptom, which was best understood as an unwitting compromise formation between the wish and the defense.

At first, however, Freud was uncertain about the precise status of the sexual component in this dynamic conception of the psyche. His patients seemed to recall actual experiences of early seductions, often incestuous in nature. Freud's initial impulse was to accept these as having happened. But then, as he disclosed in a now famous letter to Fliess of Sept. 2, 1897, he concluded that, rather than being memories of actual events, these shocking recollections were the residues of infantile impulses and desires to be seduced by an adult. What was recalled was not a genuine memory but what he would later call a screen memory, or fantasy, hiding a primitive wish. That is, rather than stressing the corrupting initiative of adults in the etiology of neuroses, Freud concluded that the fantasies and yearnings of the child were at the root of later conflict.

The absolute centrality of his change of heart in the subsequent development of psychoanalysis cannot be doubted. For in attributing sexuality to children, emphasizing the causal power of fantasies, and establishing the importance of repressed desires, Freud laid the groundwork for what many have called the epic journey into his own psyche, which followed soon after the dissolution of his partnership with Breuer.

Freud's work on hysteria had focused on female sexuality and its potential for neurotic expression. To be fully universal, psychoanalysis—a term Freud coined in 1896—would also have to examine the male psyche in a condition of what might be called normality. It would have to become more than a psychotherapy and develop into a complete theory of the mind. To this end Freud accepted the enormous risk of generalizing from the experience he knew best: his own. Significantly, his self-analysis was both the first and the last in the history of the movement he spawned; all future analysts would have to undergo a training analysis with someone whose own analysis was ultimately traceable to Freud's of his disciples.

Freud's self-exploration was apparently enabled by a disturbing event in his life. In October 1896, Jakob Freud died shortly before his 81st birthday. Emotions were released in his son that he understood as having been long repressed, emotions concerning his earliest familial experiences and feelings. Beginning in earnest in July 1897, Freud attempted to reveal their meaning by drawing on a technique that had been available for millennia: the deciphering of dreams. Freud's contribution to the tradition of dream analysis was path-breaking, for in insisting on them as "the royal road to a knowledge of the unconscious," he provided a remarkably elaborate account of why dreams originate and how they function.

The Interpretation of Dreams

In what many commentators consider his master work, *Die Traumdeutung* (published in 1899, but given the date of the dawning century to emphasize its epochal character; *The Interpretation of Dreams*), he presented his findings. Interspersing evidence from his own dreams with evidence from those recounted in his clinical practice, Freud contended that dreams played a fundamental role in the psychic economy. The mind's energy—which Freud called libido and identified principally, but not exclusively, with the sexual drive—was a fluid and malleable force capable of excessive and disturbing power. Needing to be discharged to ensure pleasure and prevent pain, it sought whatever outlet it might find. If denied the gratification provided by direct motor action, libidinal energy could seek its release through mental channels. Or, in the language of *The Interpretation of Dreams*, a wish can be satisfied by an imaginary wish fulfillment. All dreams, Freud claimed, even nightmares manifesting apparent anxiety, are the fulfillment of such wishes.

More precisely, dreams are the disguised expression of wish fulfillments. Like neurotic symptoms, they are the effects of compromises in the psyche between desires and prohibitions in conflict with their realization. Although sleep can relax the power of the mind's diurnal censorship of forbidden desires, such censorship, nonetheless, persists in part during nocturnal existence. Dreams, therefore, have to be decoded to be understood, and not merely because they are actually forbidden desires experienced in distorted fashion. For dreams undergo further revision in the process of being recounted to the analyst.

The Interpretation of Dreams provides a hermeneutic for the unmasking of the dream's disguise, or dreamwork, as Freud called it. The manifest content of the dream, that which is remembered and reported, must be understood as veiling a latent meaning. Dreams defy logical entailment and narrative coherence, for they intermingle the residues of immediate daily experience with the deepest, often most infantile wishes. Yet they can be ultimately decoded by attending to four basic activities of the dreamwork and reversing their mystifying effect.

The first of these activities, condensation, operates through the fusion of several different elements into one. As such, it exemplifies one of the key operations of psychic life, which Freud called overdetermination. No direct correspondence between a simple manifest content and its multidimensional latent counterpart can be assumed. The second activity of the dreamwork, displacement, refers to the decentring of dream thoughts, so that the most urgent wish is often obliquely or marginally represented on the manifest level. Displacement also means the associative substitution of one signifier in the dream for another, say, the king for one's father. The third activity Freud called representation, by which he meant the transformation of thoughts into images. Decoding a dream thus means translating such visual representations back into intersubjectively available language through free association. The final function of the dreamwork is secondary revision, which provides some order and intelligibility to the dream by supplementing its content with narrative coherence. The process of dream interpretation thus reverses the direction of the dreamwork, moving from the level of the conscious recounting of the dream through the preconscious back beyond censorship into the unconscious itself.

In 1904 Freud published *Zur Psychopathologie des All-*

tagslebens (*The Psychopathology of Everyday Life*), in which he explored such seemingly insignificant errors as slips of the tongue or pen (later colloquially called Freudian slips), misreadings, or forgetting of names. These errors Freud understood to have symptomatic and thus interpretable importance. But unlike dreams they need not betray a repressed infantile wish yet can arise from more immediate hostile, jealous, or egoistic causes.

In 1905 Freud extended the scope of this analysis by examining *Der Witz und seine Beziehung zum Unbewussten* (*Jokes and Their Relation to the Unconscious*). Invoking the idea of "joke-work" as a process comparable to dream-work, he also acknowledged the double-sided quality of jokes, at once consciously contrived and unconsciously revealing. Seemingly innocent phenomena like puns or jests are as open to interpretation as more obviously tendentious, obscene, or hostile jokes. The explosive response often produced by successful humour, Freud contended, owes its power to the orgasmic release of unconscious impulses, aggressive as well as sexual. But insofar as jokes are more deliberate than dreams or slips, they draw on the rational dimension of the psyche that Freud was to call the ego as much as on what he was to call the id.

Pioneering studies in sexology

In 1905 Freud also published the work that first thrust him into the limelight as the alleged champion of a pansexualist understanding of the mind: *Drei Abhandlungen zur Sexualtheorie* (*Three Contributions to the Sexual Theory,* later translated as *Three Essays on the Theory of Sexuality*), revised and expanded in subsequent editions. The work established Freud, along with Richard von Kraft-Ebing, Havelock Ellis, Albert Moll, and Iwan Bloch, as a pioneer in the serious study of sexology. Here he outlined in greater detail than before his reasons for emphasizing the sexual component in the development of both normal and pathological behaviour. Although not as reductionist as popularly assumed, Freud nonetheless extended the concept of sexuality beyond conventional usage to include a panoply of erotic impulses from the earliest childhood years on. Distinguishing between sexual aims (the act toward which instincts strive) and sexual objects (the person, organ, or physical entity eliciting attraction), he elaborated a repertoire of sexually generated behaviour of astonishing variety. Beginning very early in life, imperiously insistent on its gratification, remarkably plastic in its expression, and open to easy maldevelopment, sexuality, Freud concluded, is the prime mover in a great deal of human behaviour.

To spell out the formative development of the sexual drive, Freud focused on the progressive replacement of erotogenic zones in the body by others. An originally polymorphous sexuality first seeks gratification orally through sucking at the mother's breast, an object for which other surrogates can later be provided. Initially unable to distinguish between self and breast, the infant soon comes to appreciate its mother as the first external love object. Later Freud would contend that even before that moment, the child can treat its own body as such an object, going beyond undifferentiated autoeroticism to a narcissistic love for the self as such. After the oral phase, during the second year, the child's erotic focus shifts to its anus, stimulated by the struggle over toilet training. During the anal phase the child's pleasure in defecation is confronted with the demands of self-control. The third phase, lasting from about the fourth to the sixth year, he called the phallic. Because Freud relied on male sexuality as the norm of development, his analysis of this phase aroused considerable opposition, especially because he claimed its major concern is castration anxiety.

To grasp what Freud meant by this fear, it is necessary to understand one of his central contentions. As has been stated, the death of Freud's father was the trauma that permitted him to delve into his own psyche. Not only did Freud experience the expected grief, but he also expressed disappointment, resentment, and even hostility toward his father in the dreams he analyzed at the time. In the process of abandoning the seduction theory he recognized the source of the anger as his own psyche rather than anything objectively done by his father. Turning, as he often did, to evidence from literary and mythical texts as anticipations of his psychological insights, Freud interpreted that source in terms of Sophocles' tragedy *Oedipus Rex*. The universal applicability of its plot, he conjectured, lies in the desire of every male child to sleep with his mother and remove the obstacle to the realization of that wish, his father. What he later dubbed the Oedipus complex presents the child with a critical problem, for the unrealizable yearning at its root provokes an imagined response on the part of the father: the threat of castration.

Oedipus complex

The phallic stage can only be successfully surmounted if the Oedipus complex with its accompanying castration anxiety can be resolved. According to Freud, this resolution can occur if the boy finally suppresses his sexual desire for the mother, entering a period of so-called latency, and internalizes the reproachful prohibition of the father, making it his own with the construction of that part of the psyche Freud called the superego or the conscience.

The blatantly phallocentric bias of this account, which was supplemented by a highly controversial assumption of penis envy in the already castrated female child, proved troublesome for subsequent psychoanalytic theory. Not surprisingly, later analysts of female sexuality have paid more attention to the girl's relations with the pre-Oedipal mother than to the vicissitudes of the Oedipus complex. Anthropological challenges to the universality of the complex have also been damaging, although it has been possible to redescribe it in terms that lift it out of the specific familial dynamics of Freud's own day. If the creation of culture is understood as the institution of kinship structures based on exogamy, then the Oedipal drama reflects the deeper struggle between natural desire and cultural authority.

Freud, however, always maintained the intrapsychic importance of the Oedipus complex, whose successful resolution is the precondition for the transition through latency to the mature sexuality he called the genital phase. Here the parent of the opposite sex is conclusively abandoned in favour of a more suitable love object able to reciprocate reproductively useful passion. In the case of the girl, disappointment over the nonexistence of a penis is transcended by the rejection of her mother in favour of a father figure instead. In both cases, sexual maturity means heterosexual, procreatively inclined, genitally focused behaviour.

Sexual maladjustments

Sexual development, however, is prone to troubling maladjustments preventing this outcome if the various stages are unsuccessfully negotiated. Fixation of sexual aims or objects can occur at any particular moment, caused either by an actual trauma or the blockage of a powerful libidinal urge. If the fixation is allowed to express itself directly at a later age, the result is what was then generally called a perversion. If, however, some part of the psyche prohibits such overt expression, then, Freud contended, the repressed and censored impulse produces neurotic symptoms, neuroses being conceptualized as the negative of perversions. Neurotics repeat the desired act in repressed form, without conscious memory of its origin or the ability to confront and work it through in the present.

In addition to the neurosis of hysteria, with its conversion of affective conflicts into bodily symptoms, Freud developed complicated etiological explanations for other typical neurotic behaviour, such as obsessive-compulsions, paranoia, and narcissism. These he called psychoneuroses, because of their rootedness in childhood conflicts, as opposed to the actual neuroses such as hypochondria, neurasthenia, and anxiety neurosis, which are due to problems in the present (the last, for example, being caused by the physical suppression of sexual release).

Freud's elaboration of his therapeutic technique during these years focused on the implications of a specific element in the relationship between patient and analyst, an element whose power he first began to recognize in reflecting on Breuer's work with Anna O. Although later scholarship has cast doubt on its veracity, Freud's account of the episode was as follows. An intense rapport between Breuer and his patient had taken an alarming turn when Anna divulged her strong sexual desire for him. Breuer, who recognized the stirrings of reciprocal feelings, broke off his treatment out of an understandable confusion about the ethical implications of acting on these impulses.

Freud came to see in this troubling interaction the effects of a more pervasive phenomenon, which he called transference (or in the case of the analyst's desire for the patient, counter-transference). Produced by the projection of feelings, transference, he reasoned, is the reenactment of childhood urges cathected (invested) on a new object. As such, it is the essential tool in the analytic cure, for by bringing to the surface repressed emotions and allowing them to be examined in a clinical setting, transference can permit their being worked through in the present. That is, affective remembrance can be the antidote to neurotic repetition.

Transference

It was largely to facilitate transference that Freud developed his celebrated technique of having the patient lie on a couch, not looking directly at the analyst, and free to fantasize with as little intrusion of the analyst's real personality as possible. Restrained and neutral, the analyst functions as a screen for the displacement of early emotions, both erotic and aggressive. Transference onto the analyst is itself a kind of neurosis, but one in the service of an ultimate working through of the conflicting feelings it expresses. Only certain illnesses, however, are open to this treatment, for it demands the ability to redirect libidinal energy outward. The psychoses, Freud sadly concluded, are based on the redirection of libido back onto the patient's ego and cannot therefore be relieved by transference in the analytic situation. How successful psychoanalytic therapy has been in the treatment of psychoneuroses remains, however, a matter of considerable dispute.

Although Freud's theories were offensive to many in the Vienna of his day, they began to attract a cosmopolitan group of supporters in the early 1900s. In 1902 the Psychological Wednesday Circle began to gather in Freud's waiting room with a number of future luminaries in the psychoanalytic movements in attendance. Alfred Adler and Wilhelm Stekel were often joined by guests such as Sándor Ferenczi, Carl Gustav Jung, Otto Rank, Ernest Jones, Max Eitingon, and A.A. Brill. In 1908 the group was renamed the Vienna Psychoanalytic Society and held its first international congress in Salzburg. In the same year the first branch society was opened in Berlin. In 1909 Freud, along with Jung and Ferenczi, made a historic trip to Clark University in Worcester, Mass. The lectures he gave there were soon published as *Über Psychoanalyse* (1910; *The Origin and Development of Psychoanalysis*), the first of several introductions he wrote for a general audience. Along with a series of vivid case studies—the most famous known colloquially as "Dora" (1905), "Little Hans" (1909), "The Rat Man" (1909), "The Psychotic Dr. Schreber" (1911), and "The Wolf Man" (1918)—they made his ideas known to a wider public.

As might be expected of a movement whose treatment emphasized the power of transference and the ubiquity of Oedipal conflict, its early history is a tale rife with dissension, betrayal, apostasy, and excommunication. The most widely noted schisms occurred with Adler in 1911, Stekel in 1912, and Jung in 1913; these were followed by later breaks with Ferenczi, Rank, and Wilhelm Reich in the 1920s. Despite efforts by loyal disciples like Ernest Jones to exculpate Freud from blame, subsequent research concerning his relations with former disciples like Viktor Tausk have clouded the picture considerably. Critics of the hagiographic legend of Freud have, in fact, had a relatively easy time documenting the tension between Freud's aspirations to scientific objectivity and the extraordinarily fraught personal context in which his ideas were developed and disseminated. Even well after Freud's death, his archivists' insistence on limiting access to potentially embarrassing material in his papers has reinforced the impression that the psychoanalytic movement resembled more a sectarian church than a scientific community (at least as the latter is ideally understood).

If the troubled history of its institutionalization served to call psychoanalysis into question in certain quarters, so too did its founder's penchant for extrapolating his clinical findings into a more ambitious general theory. As he admitted to Fliess in 1900, "I am actually not a man of science at all. . . . I am nothing but a conquistador by temperament, an adventurer." Freud's so-called metapsychology soon became the basis for wide-ranging speculations about cultural, social, artistic, religious, and anthropological phenomena. Composed of a complicated and often revised mixture of economic, dynamic, and topographical elements, the metapsychology was developed in a series of 12 papers Freud composed during World War I, only some of which were published in his lifetime. Their general findings appeared in two books in the 1920s: *Jenseits des Lustprinzips* (1920; *Beyond the Pleasure Principle*) and *Das Ich und das Es* (1923; *The Ego and the Id*).

In these works, Freud attempted to clarify the relationship between his earlier topographical division of the psyche into the unconscious, preconscious, and conscious and his subsequent structural categorization into id, ego, and superego. The id was defined in terms of the most primitive urges for gratification in the infant, urges dominated by the desire for pleasure through the release of tension and the cathexis of energy. Ruled by no laws of logic, indifferent to the demands of expediency, unconstrained by the resistance of external reality, the id is ruled by what Freud called the primary process directly expressing somatically generated instincts. Through the inevitable experience of frustration the infant learns to adapt itself to the exigencies of reality. The secondary process that results leads to the growth of the ego, which follows what Freud called the reality principle in contradistinction to the pleasure principle dominating the id. Here the need to delay gratification in the service of self-preservation is slowly learned in an effort to thwart the anxiety produced by unfulfilled desires. What Freud termed defense mechanisms are developed by the ego to deal with such conflicts. Repression is the most fundamental, but Freud also posited an entire repertoire of others, including reaction formation, isolation, undoing, denial, displacement, and rationalization.

Id, ego, superego

The last component in Freud's trichotomy, the superego, develops from the internalization of society's moral commands through identification with parental dictates during the resolution of the Oedipus complex. Only partly conscious, the superego gains some of its punishing force by borrowing certain aggressive elements in the id, which are turned inward against the ego and produce feelings of guilt. But it is largely through the internalization of social norms that the superego is constituted, an acknowledgement that prevents psychoanalysis from conceptualizing the psyche in purely biologistic or individualistic terms.

Freud's understanding of the primary process underwent a crucial shift in the course of his career. Initially he counterposed a libidinal drive that seeks sexual pleasure to a self-preservation drive whose telos is survival. But in 1914, while examining the phenomenon of narcissism, he came to consider the latter instinct as merely a variant of the former. Unable to accept so monistic a drive theory, Freud sought a new dualistic alternative. He arrived at the speculative assertion that there exists in the psyche an innate, regressive drive for stasis that aims to end life's inevitable tension. This striving for rest he christened the Nirvana principle and the drive underlying it the death instinct, or Thanatos, which he could substitute for self-preservation as the contrary of the life instinct, or Eros.

Social and cultural studies. Freud's mature instinct theory is in many ways a metaphysical construct, comparable to Bergson's *élan vital* or Schopenhauer's Will. Emboldened by its formulation, Freud launched a series of audacious studies that took him well beyond his clinician's consulting room. These he had already commenced with investigations of Leonardo da Vinci (1910) and the novel *Gradiva* by Wilhelm Jensen (1907). Here Freud attempted to psychoanalyze works of art as symbolic expressions of their creator's psychodynamics.

The fundamental premise that permitted Freud to examine cultural phenomena was called sublimation in the *Three Essays*. The appreciation or creation of ideal beauty, Freud contended, is rooted in primitive sexual urges that are transfigured in culturally elevating ways. Unlike repression, which produces only neurotic symptoms whose meaning is unknown even to the sufferer, sublimation is a conflict-free resolution of repression, which leads to intersubjectively available cultural works. Although potentially reductive in its implications, the psychoanalytic

Sublimation and culture

interpretation of culture can be justly called one of the most powerful "hermeneutics of suspicion," to borrow the French philosopher Paul Ricoeur's phrase, because it debunks idealist notions of high culture as the alleged transcendence of baser concerns.

Freud extended the scope of his theories to include anthropological and social psychological speculation as well in *Totem und Tabu* (1913; *Totem and Taboo*). Drawing on Sir James Frazer's explorations of the Australian Aborigines, he interpreted the mixture of fear and reverence for the totemic animal in terms of the child's attitude toward the parent of the same sex. The Aborigines' insistence on exogamy was a complicated defense against the strong incestuous desires felt by the child for the parent of the opposite sex. Their religion was thus a phylogenetic anticipation of the ontogenetic Oedipal drama played out in modern man's psychic development. But whereas the latter was purely an intrapsychic phenomenon based on fantasies and fears, the former, Freud boldly suggested, was based on actual historical events. Freud speculated that the rebellion of sons against dominating fathers for control over women had culminated in actual parricide. Ultimately producing remorse, this violent act led to atonement through incest taboos and the prohibitions against harming the father-substitute, the totemic object or animal. When the fraternal clan replaced the patriarchal horde, true society emerged. For renunciation of individual aspirations to replace the slain father and a shared sense of guilt in the primal crime led to a contractual agreement to end internecine struggle and band together instead. The totemic ancestor then could evolve into the more impersonal God of the great religions.

A subsequent effort to explain social solidarity, *Massenpsychologie und Ich-analyse* (1921; *Group Psychology and the Analysis of the Ego*), drew on the antidemocratic crowd psychologists of the late 19th century, most notably Gustave Le Bon. Here the disillusionment with liberal, rational politics that some have seen as the seedbed of much of Freud's work was at its most explicit (the only competitor being the debunking psychobiography of Woodrow Wilson he wrote jointly with William Bullitt in 1930, which was not published until 1967). All mass phenomena, Freud suggested, are characterized by intensely regressive emotional ties stripping individuals of their self-control and independence. Rejecting possible alternative explanations such as hypnotic suggestion or imitation and unwilling to follow Jung in postulating a group mind, Freud emphasized instead individual libidinal ties to the group's leader. Group formation is like regression to a primal horde with the leader as the original father. Drawing on the army and the Roman Catholic Church as his examples, Freud never seriously considered less authoritarian modes of collective behaviour.

Freud's bleak appraisal of social and political solidarity was replicated, if in somewhat more nuanced form, in his attitude toward religion. Although many accounts of Freud's development have discerned debts to one or another aspect of his Jewish background, debts Freud himself partly acknowledged, his avowed position was deeply irreligious. As noted in the account of *Totem and Taboo,* he always attributed the belief in divinities ultimately to the displaced worship of human ancestors. One of the most potent sources of his break with former disciples like Jung was precisely this skepticism toward spirituality.

Skepticism toward religious beliefs

In his 1907 essay "Zwangshandlungen und Religionsübungen" ("Obsessive Acts and Religious Practices," later translated as "Obsessive Actions and Religious Practices") Freud had already contended that obsessional neuroses are private religious systems and religions themselves no more than the obsessional neuroses of mankind. Twenty years later, in *Die Zukunft einer Illusion* (1927; *The Future of an Illusion*), he elaborated this argument, adding that belief in God is a mythic reproduction of the universal state of infantile helplessness. Like an idealized father, God is the projection of childish wishes for an omnipotent protector. If children can outgrow their dependence, he concluded with cautious optimism, then humanity may also hope to leave behind its immature heteronomy.

The simple Enlightenment faith underlying this analysis quickly elicited critical comment, which led to its modification. In an exchange of letters with the French novelist Romain Rolland, Freud came to acknowledge a more intractable source of religious sentiment. The opening section of his next speculative tract, *Das Unbehagen in der Kultur* (1930; *Civilization and Its Discontents*), was devoted to what Rolland had dubbed the oceanic feeling. Freud described it as a sense of indissoluble oneness with the universe, which mystics in particular have celebrated as the fundamental religious experience. Its origin, Freud claimed, is nostalgia for the pre-Oedipal infant's sense of unity with its mother. Although still rooted in infantile helplessness, religion thus derives to some extent from the earliest stage of postnatal development. Regressive longings for its restoration are possibly stronger than those for a powerful father and thus cannot be worked through by way of a collective resolution of the Oedipus complex.

Civilization and Its Discontents

Civilization and Its Discontents, written after the onset of Freud's struggle with cancer of the jaw and in the midst of the rise of European Fascism, was a profoundly unconsoling book. Focusing on the prevalence of human guilt and the impossibility of achieving unalloyed happiness, Freud contended that no social solution of the discontents of mankind is possible. All civilizations, no matter how well planned, can provide only partial relief. For aggression among men is not due to unequal property relations or political injustice, which can be rectified by laws, but rather to the death instinct redirected outward.

Even Eros, Freud suggested, is not fully in harmony with civilization, for the libidinal ties creating collective solidarity are aim-inhibited and diffuse rather than directly sexual. Thus, there is likely to be tension between the urge for sexual gratification and the sublimated love for mankind. Furthermore, because Eros and Thanatos are themselves at odds, conflict and the guilt it engenders are virtually inevitable. The best to be hoped for is a life in which the repressive burdens of civilization are in rough balance with the realization of instinctual gratification and the sublimated love for mankind. But reconciliation of nature and culture is impossible, for the price of any civilization is the guilt produced by the necessary thwarting of man's instinctual drives. Although elsewhere Freud had postulated mature, heterosexual genitality and the capacity to work productively as the hallmarks of health and urged that "where id is, there shall ego be," it is clear that he held out no hope for any collective relief from the discontents of civilization. He only offered an ethic of resigned authenticity, which taught the wisdom of living without the possibility of redemption, either religious or secular.

Freud's final major work, *Der Mann Moses und die monotheistische Religion* (1938; *Moses and Monotheism*), was more than just the "historical novel" he had initially thought to subtitle it. Moses had long been a figure of capital importance for Freud; indeed Michelangelo's famous statue of Moses had been the subject of an essay written in 1914. The book itself sought to solve the mystery of Moses' origins by claiming that he was actually an aristocratic Egyptian by birth who had chosen the Jewish people to keep alive an earlier monotheistic religion. Too stern and demanding a taskmaster, Moses was slain in a Jewish revolt, and a second, more pliant leader, also called Moses, rose in his place. The guilt engendered by the parricidal act was, however, too much to endure, and the Jews ultimately returned to the religion given them by the original Moses as the two figures were merged into one in their memories. Here Freud's ambivalence about his religious roots and his father's authority was allowed to pervade a highly fanciful story that reveals more about its author than its ostensible subject.

Moses and Monotheism

Moses and Monotheism was published in the year Hitler invaded Austria. Freud was forced to flee to England. His books were among the first to be burned, as the fruits of a "Jewish science," when the Nazis took over Germany. Although psychotherapy was not banned in the Third Reich, where Field Marshall Hermann Göring's cousin headed an official institute, psychoanalysis essentially went into exile, most notably to North America and England. Freud himself died only a few weeks after World War II broke out, on Sept. 23, 1939, at a time when his worst fears about the irrationality lurking behind the facade of civilization were being realized. Freud's death did not, however, hinder the reception and dissemination of his ideas. A plethora of Freudian schools emerged to develop psychoanalysis in different directions. In fact, despite the relentless and often compelling challenges mounted against virtually all of his ideas, Freud has remained one of the most potent figures in the intellectual landscape of the 20th century.

(M.E.J.)

MAJOR WORKS. Several of these works appeared originally as journal articles; only the publication in book form is cited here. *Studien über Hysterie*, with Josef Breuer (1895; *Studies in Hysteria*, 1936); *Die Traumdeutung* (1899, dated 1900; *The Interpretation of Dreams*, 1913); *Zur Psychopathologie des Alltagslebens* (1904; *Psychopathology of Everyday Life*, 1914); *Drei Abhandlungen zur Sexualtheorie* (1905; *Three Contributions to the Sexual Theory*, 1910); *Über Psychoanalyse* (1910; *The Origin and Development of Psychoanalysis*, 1949); *Totem und Tabu: einige Übereinstimmungen im Seelenleben der Wilden und der Neurotiker* (1913; *Totem and Taboo: Resemblances Between the Psychic Lives of Savages and Neurotics*, 1918); *Zur Geschichte der psychoanalytischen Bewegung* (1924; *The History of the Psychoanalytic Movement*, 1917); *Vorlesungen zur Einführung in die Psychoanalyse* (1916–17; *A General Introduction to Psychoanalysis*, 1920); *Jenseits des Lustprinzips* (1920; *Beyond the Pleasure Principle*, 1922); *Das Ich und das Es* (1923; *The Ego and the Id*, 1927); *Hemmung, Symptom und Angst* (1926; *Inhibition, Symptoms and Anxiety*, 1927); *Die Frage der Laienanalyse* (1926; *The Problem of Lay-Analyses*, 1927); *Die Zukunft einer Illusion* (1927; *The Future of an Illusion*, 1928); *Das Unbehagen in der Kultur* (1930; *Civilization and Its Discontents*, 1930); *Neue Folge der Vorlesungen zur Einführung in die Psychoanalyse* (1933; *New Introductory Lectures on Psycho-Analysis*, 1933); *Der Mann Moses und die monotheistische Religion* (1939; *Moses and Monotheism*, 1939).

The standard German edition of Freud's works is *Gesammelte Werke: Chronologisch geordnet*, 18 vol. in 17 (1940–68). The English edition, with better annotations than the original, is *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. trans. and ed. by JAMES STRACHEY, *et al.*, 24 vol. (1953–74, reprinted 1981); it is complemented by SAMUEL A. GUTTMAN, STEPHEN M. PARRISH, and RANDALL L. JONES (eds.), *The Concordance to The Standard Edition of the Complete Psychological Works of Sigmund Freud*, 6 vol., 2nd ed. (1984). Also helpful is ALEXANDER GRINSTEIN (comp.), *Sigmund Freud's Writings: A Comprehensive Bibliography* (1977), including listings of works not found in *The Standard Edition* and an index of English titles of Freud's works.

BIBLIOGRAPHY. Among biographical works are SIGMUND FREUD, *An Autobiographical Study*, 2nd ed. (1946, reissued 1963; originally published in German, 1925), his own brief account of his career and theories; ERNEST JONES, *The Life and Work of Sigmund Freud*, 3 vol. (1953–57, reissued 1981; also published as *Sigmund Freud: Life and Work*, 1953–57), also available in a one-volume condensed edition with the same title, edited and abridged by LIONEL TRILLING and STEVEN MARCUS (1961, reissued 1964); RICHARD WOLLHEIM, *Sigmund Freud* (1971, reissued 1981); PHILIP RIEFF, *Freud: The Mind of a Moralist*, 3rd ed. (1979); RONALD W. CLARK, *Freud. The Man and the Cause* (1980); and PETER GAY, *Freud: A Life for Our Time* (1980).

Selections from Freud's original writings and correspondence include JEFFREY MOUSSAIEFF MASSON (trans. and ed.), *The Complete Letters of Sigmund Freud to Wilhelm Fliess, 1877–1904* (1985); ERNST L. FREUD (ed.), *Letters of Sigmund Freud, 1873–1939* (1961, reprinted 1975; originally published in German, 1960); HILDA C. ABRAHAM and ERNST L. FREUD (eds.), *A Psycho-Analytic Dialogue: The Letters of Sigmund Freud and Karl Abraham, 1907–1926* (1965; originally published in German, 1965); ERNST L. FREUD (ed.), *The Letters of Sigmund Freud and Arnold Zweig* (1970, reprinted 1987; originally published in German, 1968); NATHAN G. HALE, JR. (ed.), *James Jackson Putnam and Psychoanalysis: Letters Between Putnam and Sigmund Freud, Ernest Jones, William James, Sandor Ferenczi, and Morton Prince, 1877–1917* (1971); ERNST PFEIFFER (ed.), *Sigmund Freud and Lou Andreas-Salomé: Letters* (1972, reissued 1985; originally published in German, 1966); WILLIAM MCGUIRE (ed.), *The Freud/Jung Letters*, trans. from German (1974, reprinted 1979); R. ANDREWS PASKAUSKAS (ed.), *The Complete Correspondence of Sigmund Freud and Ernest Jones, 1908–1939* (1993); and EVA BRABANT, ERNST FALZADER, and PATRIZIA GIAMPIERI-DEUTSCH (eds.), *The Correspondence of Sigmund Freud and Sándor Ferenczi* (1993–   ).

Views by Freud's family, friends, and colleagues include FRITZ WITTELS, *Sigmund Freud: His Personality, His Teaching, & His School* (1924, reprinted 1971; originally published in German, 1924); THEODOR REIK, *From Thirty Years with Freud*, trans. from German (1940, reissued 1975); HANNS SACHS, *Freud* (1944, reissued 1970); MARTIN FREUD, *Glory Reflected: Sigmund Freud, Man and Father* (1957; also published as *Sigmund Freud: Man and Father*, 1928, reissued 1983), by one of his children; ERICH FROMM, *Sigmund Freud's Mission: An Analysis of His Personality and Influence* (1959, reprinted 1978); MARY HIGGINS and CHESTER M. RAPHAEL (eds.), *Reich Speaks of Freud: Wilhelm Reich Discusses His Work and His Relationship with Sigmund Freud* (1967, reissued 1975); MAX SCHUR, *Freud* (1972); and ALDO CAROTENUTO, *A Secret Symmetry: Sabina Spielrein Between Jung and Freud* (1982; originally published in Italian, 1980).

Contemporaries and associates are described in VINCENT BROME, *Freud and His Early Circle: The Struggles of Psycho-Analysis* (1967); and PAUL ROAZEN, *Freud and His Followers* (1975, reissued 1984), and *Brother Animal: The Story of Freud and Tausk* (1969, reprinted 1986). Also of interest is K.R. EISSLER, *Talent and Genius: The Fictitious Case of Tausk Contra Freud* (1971).

Histories of psychoanalysis and psychoanalytic theory are offered in MARIE JAHODA, *Freud and the Dilemmas of Psychology* (1977, reissued 1981); SEYMOUR FISHER and ROGER P. GREENBERG, *The Scientific Credibility of Freud's Theories and Therapy* (1977, reprinted 1985), and *The Scientific Evaluation of Freud's Theories and Therapy: A Book of Readings* (1977); FRANK J. SULLOWAY, *Freud, Biologist of the Mind* (1979, reprinted 1983); ALEXANDER GRINSTEIN, *Sigmund Freud's Dreams*, 2nd ed. (1980); BRUNO BETTELHEIM, *Freud and Man's Soul* (1983); MARSHALL EDELSON, *Hypothesis and Evidence in Psychoanalysis* (1984); and WILLIAM J. MCGRATH, *Freud's Discovery of Psychoanalysis: The Politics of Hysteria* (1986).

Interpretive studies of Freud's work and views include HERBERT MARCUSE, *Eros and Civilisation: A Philosophical Inquiry into Freud* (1955, reissued 1974); J.A.C. BROWN, *Freud and the Post-Freudians* (1961, reprinted 1985); PHILIP RIEFF, *The Triumph of the Therapeutic: Uses of Faith After Freud* (1966, reissued 1987); PAUL ROAZEN, *Freud: Political and Social Thought* (1968, reissued 1986); PAUL RICOEUR, *Freud and Philosophy* (1970, originally published in French, 1961); and JULIET MITCHELL, *Psychoanalysis and Feminism* (1974).

Recent critiques of Freudian theory include ERICH FROMM, *Greatness and Limitations of Freud's Thought* (1980); JANET MALCOLM, *In the Freud Archives* (1984); JEFFREY MOUSSAIEFF MASSON, *The Assault on Truth: Freud's Suppression of the Seduction Theory* (1984, reissued 1994; also published as *Freud: The Assault on Truth*, 1984); ADOLF GRÜNBAUM, *The Foundations of Psychoanalysis* (1984); and ROBERT R. HOLT, *Freud Reappraised: A Fresh Look at Psychoanalytic Theory* (1989). PAUL ROBINSON, *Freud and His Critics* (1993), is a defense against several critics.

Freud's major case studies are reappraised in KARIN OBHOLZER, *The Wolf-Man: Conversations with Freud's Patient—Sixty Years Later* (1982; originally published in German, 1980); PATRICK J. MAHONY, *Freud and the Rat Man* (1986); FRANK J. SULLOWAY, "Reassessing Freud's Case Histories: The Social Construction of Psychoanalysis," *Isis*, 82:245–275 (1991, reprinted in TOBY GELFAND and JOHN KERR [eds.], *Freud and the History of Psychoanalysis*, 1992); HANNAH S. DECKER, *Freud, Dora, and Vienna 1900* (1991); and ROGIN TOLMACH LAKOFF and JAMES C. COYNE, *Father Knows Best: The Use and Abuse of Power in Freud's Case of Dora* (1993). (M.E.J./Ed.)

Fossil Fuels

Fossil fuels are combustible materials of organic origin, including oil, bitumen, natural gas, and coal. Such materials were formed by geochemical processes from the remains of organisms that were buried in the geologic past. Since the late 18th century, these mineral fuels have become the primary sources of energy for the industrial nations of the world. Together, the fossil fuels supply almost 90 percent of all the energy consumed today.

Recent scientific assessments place the world's total original endowment of recoverable conventional (*i.e.*, light and medium) oil at approximately 2,390,000,000,000 barrels. Thirty percent of this oil, however, has already been consumed, and only 23 percent is estimated as still undiscovered. Seventy-eight percent of the unconsumed conventional oil is thought to be located in the Eastern Hemisphere.

The world's total original endowment of recoverable natural gas has been assessed as the energy equivalent of about 2,576,000,000,000 barrels of oil. This includes an estimated 463,000,000,000 barrels of natural gas liquids (propane, butane, and natural gasoline), a figure based on average gas liquids production. Approximately 55 percent of the world's natural gas is thought to have been already discovered and about 14 percent consumed. Of the remaining existing gas, 84 percent is estimated to be located in the Eastern Hemisphere.

Major deposits of heavy oils are known to lie in the Western Hemisphere. The world's original endowment of recoverable heavy oil is believed to have been roughly 467,500,000,000 barrels, approximately 80 percent of which has been discovered and only about 14 percent of which has been consumed. Fifty-two percent of the unconsumed heavy oil is located in the Western Hemisphere.

Known bitumen deposits have been projected to contain about 354,000,000,000 barrels of recoverable crude oil. Seventy-six percent of the bitumen resource is in the Western Hemisphere. Known oil shale deposits have been estimated to contain about 1,067,000,000,000 barrels of recoverable shale oil. Eighty-eight percent of this resource is in the Western Hemisphere.

In summary, the world's total original endowment of recoverable petroleum, including natural gas, is assessed as equaling about 6,853,000,000,000 barrels of oil. The lighter, more desirable petroleum fuels occur primarily in the Eastern Hemisphere, while most of the heavier, less desirable varieties are found in the Western Hemisphere. The heavier fuels have not been exploited as extensively as the lighter fuels, because they are much more costly and difficult to extract and must be produced at considerably slower rates. Consequently, they remain available for development as the lighter fuels become depleted. Their cost, however, will be high and will be reflected in the world economy.

As for the other major fossil fuel, coal, scientific studies place the world's original endowment of minable coal deposits at more than 7,600,000,000,000 metric tons (approximately 6,900,000,000,000 short tons). Of this total, only about 2.5 percent has so far been exploited; therefore, coal reserves are projected to last from a few hundred to more than 1,000 years, depending on the assumed rate of consumption and technological developments. Almost 60 percent of the world's remaining recoverable coal is held by only three countries—the United States, Russia, and China.

This article describes the properties and origin of the fossil fuels and treats their occurrence, exploitation, and world distribution. For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 214, and the *Index*. For specific information about the methods of fossil fuel extraction and production, see the *Macropædia* article INDUSTRIES, EXTRACTION AND PROCESSING. (J.P.Ri./Ed.)

The article is divided into the following sections:

Oil

Liquid and gaseous hydrocarbons are so intimately associated in nature that it has become customary to shorten the expression "petroleum and natural gas" to "petroleum" when referring to both. The word petroleum (literally "rock oil" from the Latin *petra*, "rock" or "stone," and *oleum*, "oil") was first used in 1556 in a treatise published by the German mineralogist Georg Bauer, known as Georgius Agricola.

HISTORY OF USE

Exploitation of surface seeps. Small surface occurrences of petroleum in the form of natural gas and oil seeps have been known from early times. The ancient Sumerians, Assyrians, and Babylonians used crude oil and asphalt ("pitch") collected from large seeps at Tuttul (modern-day Hit) on the Euphrates for many purposes more than 5,000 years ago. Liquid oil was first used as a medicine by the ancient Egyptians, presumably as a wound dressing, liniment, and laxative.

Oil products were valued as weapons of war in the ancient world. The Persians used incendiary arrows wrapped in oil-soaked fibres at the siege of Athens in 480 BC. Early in the Christian era the Arabs and Persians distilled crude oil to obtain flammable products for military purposes. Probably as a result of the Arab invasion of Spain, the industrial art of distillation into illuminants became available in western Europe by the 12th century.

Early distillation

Several centuries later, Spanish explorers discovered oil seeps in present-day Cuba, Mexico, Bolivia, and Peru. In North America oil seeps were plentiful and were noted by early explorers in what are now New York and Pennsylvania, where the Indians were reported to have used the oil for medicinal purposes.

Extraction from underground reservoirs. Until the beginning of the 19th century, illumination in the United States and in many other countries was little improved over that known by the early Greeks and Romans. The need for better illumination that accompanied the increasing development of urban centres made it necessary to search for new sources of oil, especially since whales, which had long provided fuel for lamps, were becoming harder and harder to find. By the mid-19th century kerosene, or coal oil, derived from coal was in common use in both North America and Europe.

The Industrial Revolution brought on an ever-growing demand for a cheaper and more convenient source of lubricants as well as illuminating oil. It also required better sources of energy. Energy had previously been provided by human and animal muscle and later by the combustion of such solid fuels as wood, peat, and coal. These were collected with considerable effort and laboriously transported to the site where the energy source was needed. Liquid petroleum, on the other hand, was a more easily transportable source of energy. Oil was a much more concentrated and flexible form of fuel than anything previously available.

The stage was set for the first well specifically drilled for oil, a project undertaken by Edwin L. Drake in northwestern Pennsylvania. The completion of the well in August 1859 established the groundwork for the petroleum industry and ushered in the closely associated modern industrial age. Within a short time inexpensive oil from underground reservoirs was being processed at already existing coal-oil refineries, and by the end of the century oil fields had been discovered in 14 states from New York to California and from Wyoming to Texas. During the same period, oil fields were found in Europe and East Asia as well.

Significance of oil in modern times. At the beginning of the 20th century the Industrial Revolution had progressed to the extent that the use of refined oil for illuminants ceased to be of primary importance. The oil industry became the major supplier of energy largely because of the advent of the automobile. Although oil constitutes a major petrochemical feedstock, its primary importance is as an energy source on which the world economy depends.

The significance of oil as a world energy source is difficult to overdramatize. The growth in energy production during the 20th century is unprecedented, and increasing oil production has been by far the major contributor to that growth. Every day an immense and intricate system moves more than 60,000,000 barrels of oil from producers to consumers. The production and consumption of oil is of vital importance to international relations and has frequently been a decisive factor in the determination of foreign policy. The position of a country in this system depends on its production capacity as related to its consumption. The possession of oil deposits is sometimes the determining factor between a rich and a poor country. For any country, however, the presence or absence of oil has a major economic consequence.

Current daily consumption

On a time scale within the span of prospective human history, the utilization of oil as a major source of energy will be a transitory affair of about 100 years. Nonetheless, it will have been an affair of profound importance to world industrialization.

PROPERTIES

Chemical composition. *Hydrocarbon content.* Although oil consists basically of compounds of only two elements, carbon and hydrogen, these elements form a large variety of complex molecular structures. Regardless of physical or chemical variations, however, almost all crude oil ranges from 82 to 87 percent carbon by weight and 12 to 15 percent hydrogen. The more viscous bitumens generally vary from 80 to 85 percent carbon and from 8 to 11 percent hydrogen.

Crude oil can be grouped into three basic chemical series: paraffins, naphthenes, and aromatics. Most crude oils are mixtures of these three series in various and seemingly endless proportions. No two crude oils from different sources are completely identical.

The paraffin series of hydrocarbons, also called the methane (CH_4) series, comprises the most common hydrocarbons in crude oil. It is a saturated straight-chain series that has the general formula C_nH_{2n+2}, in which C is carbon, H is hydrogen, and n is an integer. The paraffins that are liquid at normal temperatures but boil between 40° and 200° C (approximately between 100° and 400° F) are the major constituents of gasoline. The residues obtained by refining lower-density paraffins are both plastic and solid paraffin waxes.

The naphthene series has the general formula C_nH_{2n} and is a saturated closed-ring series. This series is an important part of all liquid refinery products, but it also forms most of the complex residues from the higher boiling-point ranges. For this reason, the series is generally heavier. The residue of the refinery process is an asphalt, and the crude oils in which this series predominates are called asphalt-base crudes.

The aromatic series has the general formula C_nH_{2n-6} and is an unsaturated closed-ring series. Its most common member, benzene (C_6H_6), is present in all crude oils, but the aromatics as a series generally constitute only a small percentage of most crudes.

Nonhydrocarbon content. In addition to the practically infinite mixtures of hydrocarbon compounds that form crude oil, sulfur, nitrogen, and oxygen are usually present in small but often important quantities. Sulfur is the third most abundant atomic constituent of crude oils. It is present in the medium and heavy fractions of crude oils. In the low and medium molecular ranges, sulfur is associated only with carbon and hydrogen, while in the heavier fractions it is frequently incorporated in the large polycyclic molecules that also contain nitrogen and oxygen. The total sulfur in crude oil varies from below 0.05 percent (by weight), as in some Pennsylvania oils, to about 2 percent for average Middle Eastern crudes and up to 5 percent or more in heavy Mexican or Mississippi oils. Generally, the higher the specific gravity of the crude oil, the greater is its sulfur content. The excess sulfur is removed from crude oil during refining, because sulfur oxides released into the atmosphere during the combustion of oil would constitute a major pollutant.

Sulfur content

The oxygen content of crude oil is usually less than 2 percent by weight and is present as part of the heavier hydrocarbon compounds in most cases. For this reason, the heavier oils contain the most oxygen. Nitrogen is present in almost all crude oils, usually in quantities of less than 0.1 percent by weight. Sodium chloride also occurs in most crudes and is usually removed like sulfur.

Many metallic elements are found in crude oils, including most of those that occur in seawater. This is probably due to the close association between seawater and the organic forms from which oil is generated. Among the most common metallic elements in oil are vanadium and nickel, which apparently occur in organic combinations as they do in living plants and animals.

Crude oil also may contain a small amount of decay-resistant organic remains, such as siliceous skeletal fragments, wood, spores, resins, coal, and various other remnants of former life.

Physical properties. Oil consists of a closely related series of complex hydrocarbon compounds that range from gasoline to heavy solids. The various mixtures that constitute crude oil can be separated by distillation under increasing temperatures into such components as (from light to heavy) gasoline, kerosene, gas oil, lubricating oil, residual fuel oil, bitumen, and paraffin.

Crude oils vary greatly in their chemical composition. Because they consist of mixtures of thousands of hydrocarbon compounds, their physical properties such as colour, specific gravity, and viscosity also vary widely.

Specific gravity. Crude oil is immiscible with and lighter than water; hence it floats. Crude oils are generally classified as tar sands, heavy oils, and medium and light oils on the basis of specific gravity (*i.e.,* the ratio of the weight of equal volumes of the oil and pure water at standard conditions, with pure water considered to equal 1) and relative mobility. Tar sands contain immobile oil, which does not flow into a well bore (see below). Heavy crude oils have enough mobility that, given time, they can be obtained through a well bore in response to enhanced recovery methods. The more mobile medium and light oils are recoverable through production wells.

The API gravity scale

The widely used American Petroleum Institute (API) gravity scale is based on pure water, with an arbitrarily assigned API gravity of 10°. Liquids lighter than water, such as oil, have API gravities numerically greater than 10. Crude oils below 20° API gravity are usually considered heavy, whereas the conventional crudes with API gravities between 20° and 25° are regarded as medium, with light oils ranging above 25°.

Boiling and freezing points. Because oil is always at a temperature above the boiling point of some of its compounds, the more volatile constituents constantly escape into the atmosphere unless confined. It is impossible to refer to a common boiling point for crude oil because of the widely differing boiling points of its numerous compounds, some of which may boil at temperatures too high to be measured.

By the same token, it is impossible to refer to a common freezing point for a crude oil because the individual compounds solidify at different temperatures. However, the pour point—the temperature below which crude oil becomes plastic and will not flow—is important to recovery and transport and is always determined. Pour points range from 32° C to below −57° C.

Measurement systems. In the United States, crude oil is measured in barrels of 42 gallons each; the weight per barrel of API 30° light oil would be about 306 pounds. In many other countries, crude oil is measured in metric tons. For oil having the same gravity, a metric ton is equal to approximately 252 imperial gallons or about 7.2 U.S. barrels.

ORIGIN

Formation process. *From planktonic remains to kerogen.* Although it is recognized that the original source of carbon and hydrogen was in the materials that made up the primordial Earth, it is generally accepted that these two elements have had to pass through an organic phase to be combined into the varied complex molecules recognized as crude oil. The organic material that is the source of most oil has probably been derived from single-celled planktonic (free-floating) plants, such as diatoms and blue-green algae, and single-celled planktonic animals, such as foraminifera, which live in aquatic environments of marine, brackish, or fresh water. Such simple organisms are known to have been abundant long before the Paleozoic Era, which began some 540,000,000 years ago.

Rapid burial of the remains of the single-celled planktonic plants and animals within fine-grained sediments effectively preserved them. This provided the organic materials, the so-called protopetroleum, for later diagenesis (*i.e.,* the series of processes involving biological, chemical, and physical changes) into true petroleum.

The first, or immature, stage of petroleum formation is dominated by biological activity and chemical rearrangement, which convert organic matter to kerogen. This dark-coloured, insoluble product of bacterially altered plant and animal detritus is the source of most hydrocarbons generated in the later stages. During the first stage, biogenic methane is the only hydrocarbon generated in commercial quantities. The production of biogenic methane gas is part of the process of decomposition of organic matter carried out by anaerobic microorganisms (those capable of living in the absence of free oxygen).

From kerogen to petroleum. Deeper burial by continuing sedimentation, increasing temperatures, and advancing geologic age result in the mature stage of petroleum formation, during which the full range of petroleum compounds is produced from kerogen and other precursors by thermal degradation and cracking (the process by which heavy hydrocarbon molecules are broken up into lighter molecules). Depending on the amount and type of organic matter, oil generation occurs during the mature stage at depths of about 760 to 4,880 metres (2,500 to 16,000 feet) at temperatures between 65° and 150° C. This special environment is called the "oil window." In areas of higher than normal geothermal gradient (increase in temperature with depth), the oil window exists at shallower depths in younger sediments but is narrower. Maximum oil generation occurs from depths of 2,000 to 2,900 metres. Below 2,900 metres primarily wet gas, a type of gas containing liquid hydrocarbons known as natural gas liquids, is formed.

The "oil window"

Approximately 90 percent of the organic material in sedimentary source rocks is dispersed kerogen. Its composition varies, consisting as it does of a range of residual materials whose basic molecular structure takes the form of stacked sheets of aromatic hydrocarbon rings in which atoms of sulfur, oxygen, and nitrogen also occur. Attached to the ends of the rings are various hydrocarbon compounds, including normal paraffin chains. The mild heating of the kerogen in the oil window of a source rock over long periods of time results in the cracking of the kerogen molecules and the release of the attached paraffin chains. Further heating, perhaps assisted by the catalytic effect of clay minerals in the source rock matrix, may then produce soluble bitumen compounds, followed by the various saturated and unsaturated hydrocarbons, asphaltenes, and others of the thousands of hydrocarbon compounds that make up crude oil mixtures.

At the end of the mature stage, below about 4,880 metres, depending on the geothermal gradient, kerogen becomes condensed in structure and chemically stable. In this environment, crude oil is no longer stable and the main hydrocarbon product is dry thermal methane gas.

The geologic environment. *Origin in source beds.* Knowing the maximum temperature reached by a potential source rock during its geologic history helps in estimating the maturity of the organic material contained within it. Also, this information may indicate whether a region is gas-prone, oil-prone, both, or neither. The techniques employed to assess the maturity of potential source rocks in core samples include measuring the degree of darkening of fossil pollen grains and the colour changes in conodont fossils. In addition, geochemical evaluations can be made of mineralogical changes that were also induced by fluctuating paleotemperatures. In general, there appears to be a progressive evolution of crude oil characteristics from geologically younger, heavier, darker, more aromatic crudes to older, lighter, paler, more paraffinic types. There are, however, many exceptions to this rule, especially in regions with high geothermal gradients.

Accumulations of petroleum are usually found in relatively coarse-grained, permeable, and porous sedimentary reservoir rocks that contain little, if any, insoluble organic matter. It is unlikely that the vast quantities of oil now present in some reservoir rocks could have been generated from material of which no trace remains. Therefore, the site where commercial amounts of oil originated apparently is not always identical to the location at which they are ultimately discovered.

Primary migration

Oil is believed to have been generated in significant volumes only in fine-grained sedimentary rocks (usually clays, shales, or clastic carbonates) by geothermal action on kerogen, leaving an insoluble organic residue in the source rock. The release of oil from the solid particles of kerogen and its movement in the narrow pores and capillaries of the source rock is termed primary migration.

Accumulating sediments can provide energy to the migration system. Primary migration may be initiated during compaction as a result of the pressure of overlying sediments. Continued burial causes clay to become dehydrated by the removal of water molecules that were loosely combined with the clay minerals. With increasing temperature, the newly generated hydrocarbons may become sufficiently mobile to leave the source beds in solution, suspension, or emulsion with the water being expelled from the compacting molecular lattices of the clay minerals. The hydrocarbon molecules would compose only a very small part of the migrating fluids, a few hundred parts per million.

Migration through carrier beds. The hydrocarbons expelled from a source bed next move through the wider pores of carrier beds (*e.g.*, sandstones or carbonates) that are coarser-grained and more permeable. This movement is termed secondary migration. The distinction between primary and secondary migration is based on pore size and rock type. In some cases, oil may migrate through such permeable carrier beds until it is trapped by a permeability barrier and forms an oil accumulation (Figure 1). In others, the oil may continue its migration until it becomes a seep on the surface of the Earth, where it will be broken down chemically by oxidation and bacterial action.

Since nearly all pores in subsurface sedimentary formations are water-saturated, the migration of oil takes place in an aqueous environment. Secondary migration may result from active water movement or can occur independently, either by displacement or by diffusion. Because the specific gravity of the water in the sedimentary formation is considerably higher than that of oil, the oil will float to the surface of the water in the course of geologic time and accumulate in the highest portion of a trap.

Accumulation in reservoir beds. The porosity (volume of pore spaces) and permeability (capacity for transmitting fluids) of carrier and reservoir beds are important factors in the migration and accumulation of oil. Most petroleum accumulations have been found in clastic reservoirs (sandstones and siltstones). Next in number are the carbonate reservoirs (limestones and dolomites). Accumulations of petroleum also occur in shales and igneous and metamorphic rocks because of porosity resulting from fracturing, but such reservoirs are relatively rare. Porosities in reservoir rocks usually range from about 5 to 30 percent, but all available pore space is not occupied by petroleum. A certain amount of residual formation water cannot be displaced and is always present.

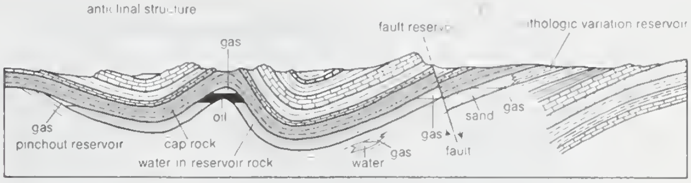


Figure 1: Principal types of traps.

Principal types of reservoir rocks

Reservoir rocks may be divided into two main types: (1) those in which the porosity and permeability is primary, or inherent, and (2) those in which they are secondary, or induced. Primary porosity and permeability are dependent on the size, shape, and grading and packing of the sediment grains (Figure 2) and also on the manner of their initial consolidation. Secondary porosity and permeability result from postdepositional factors, such as solution, recrystallization, fracturing, weathering during temporary exposure at the Earth's surface, and further cementation. These secondary factors may either enhance or diminish the inherent conditions.

Oil traps. After secondary migration in carrier beds, oil finally collects in a trap. The fundamental characteristic of a trap is an upward convex form of porous and permeable reservoir rock that is sealed above by a denser, relatively impermeable cap rock (*e.g.*, shale or evaporites). The trap may be of any shape, the critical factor being that it is a closed, inverted container. A rare exception is hydrodynamic trapping, in which high water saturation of low-permeability sediments reduces hydrocarbon permeability to near zero, resulting in a water block and an accumulation of petroleum down the structural dip of a sedimentary bed below the water in the sedimentary formation.

Anticlines

Structural traps. Traps can be formed in many ways (Figure 1). Those formed by tectonic events, such as folding or faulting of rock units, are called structural traps. The most common structural traps are anticlines, upfolds of strata that appear as ovals on the horizontal planes of geologic maps. About 80 percent of the world's petroleum has been found in anticlinal traps. Most anticlines were produced by lateral pressure, but some have resulted from the draping and subsequent compaction of accumulating sediments over topographic highs. The closure of an anticline is the vertical distance between its highest point and the spill plane, the level at which the petroleum can escape if the trap is filled beyond capacity. Some traps are filled with petroleum to their spill plane, but others contain considerably smaller amounts than they can accommodate on the basis of their size.

Another kind of structural trap is the fault trap. Here, rock fracture results in a relative displacement of strata

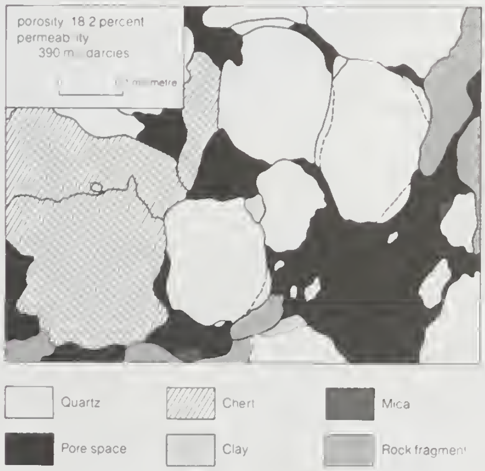


Figure 2: An example of primary porosity in a sandstone (graywacke) of an excellent reservoir rock.

Table 1: The Recoverable Oil Resources of the World*

region	cumulative production	reserves	undiscovered resources	total oil endowment†
North America	202	106	121	429
South America	74	93	44	211
Western Europe	23	19	28	70
Eastern Europe (including Russia)	113	104	64	281
Central Asia and Transcaucasia	16	24	39	79
Middle East	194	666	122	982
Africa (including North Africa)	57	62	48	167
Oceania and Asia	45	45	81	171
Total world	724	1,119	547	2,390

*In billion barrels; figures adapted from *Oil & Gas Journal* and U.S. Geological Survey.
†Percent of original reserves by average API gravity: 10°—20°: 5 percent; 20°—25°: 6 percent; 25°—35°: 57 percent; above 35°: 32 percent.

that forms a barrier to petroleum migration. A barrier can occur when an impermeable bed is brought into contact with a carrier bed. Sometimes the faults themselves provide a seal against "updip" migration when they contain impervious clay gouge material between their walls. Faults and folds often combine to produce traps, each providing a part of the container for the enclosed petroleum. Faults can, however, allow the escape of petroleum from a former trap if they breach the cap rock seal.

Other structural traps are associated with salt domes. Such traps are formed by the upward movement of salt masses from deeply buried evaporite beds, and they occur along the folded or faulted flanks of the salt plug or on top of the plug in the overlying folded or draped sediments.

Stratigraphic traps. A second major class of oil traps is the stratigraphic trap. It is related to sediment deposition or erosion and is bounded on one or more sides by zones of low permeability. Because tectonics ultimately control deposition and erosion, however, few stratigraphic traps are completely without structural influence. The geologic history of most sedimentary basins contains the prerequisites for the formation of stratigraphic traps. Typical examples are fossil carbonate reefs, marine sandstone bars, and deltaic distributary channel sandstones. When buried, each of these geomorphic features provides a potential reservoir, which is often surrounded by finer-grained sediments that may act as source or cap rocks.

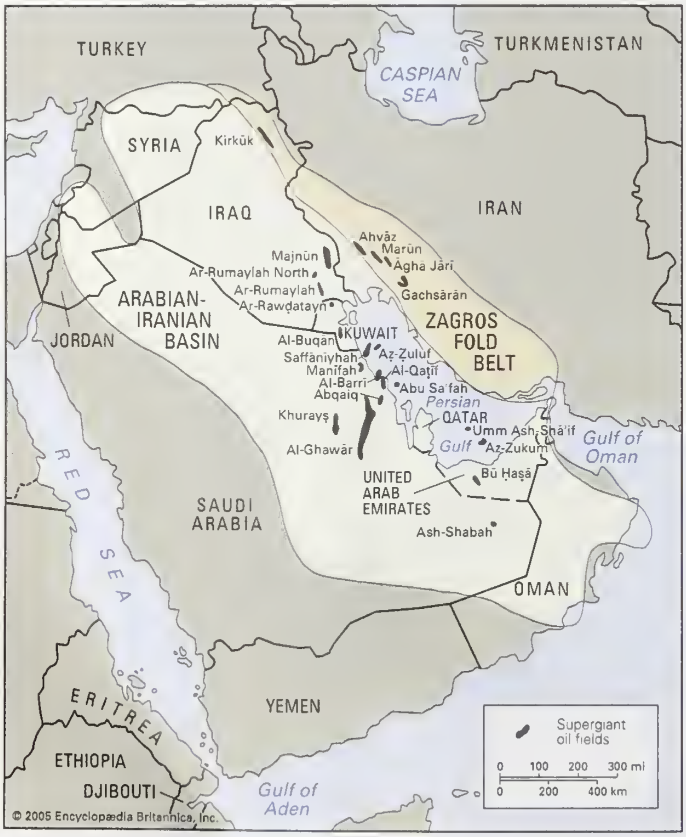


Figure 3: Major oil fields of the Arabian-Iranian basin region.

Sediments eroded from a landmass and deposited in an adjacent sea change from coarse- to fine-grained with increasing depth of water and distance from shore. Permeable sediments thus grade into impermeable sediments, forming a permeability barrier that eventually could trap migrating petroleum.

There are many other types of stratigraphic traps. Some are associated with the many transgressions and regressions of the sea that have occurred over geologic time and the resulting deposits of differing porosities. Others are caused by processes that increase secondary porosity, such as the dolomitization of limestones or the weathering of strata once located at the Earth's surface.

WORLD DISTRIBUTION

Location of reserves. *Oil fields.* Two overriding principles apply to world petroleum production. First, most petroleum is contained in a few large fields, but most fields are small. Second, as exploration progresses, the average size of the fields discovered decreases, as does the amount of petroleum found per unit of exploratory drilling. In any region, the large fields are usually discovered first.

Supergiant and world-class giant oil fields

Since exploration for oil began during the early 1860s, some 50,000 oil fields have been discovered. More than 90 percent of these fields are insignificant in their impact on world oil production. The two largest classes of fields are the supergiants, fields with 5,000,000,000 or more barrels of ultimately recoverable oil, and world-class giants, fields with 500,000,000 to 5,000,000,000 barrels of ultimately recoverable oil. Fewer than 40 supergiant oil fields have been found worldwide, yet these fields originally contained about one-half of all the oil so far discovered. The Arabian-Iranian sedimentary basin in the Persian Gulf region contains two-thirds of these supergiant fields. The remaining supergiants are distributed as follows: two in the United States, two in Russia, two in Mexico, one in Libya, one in Algeria, one in Venezuela, and two in China.

The nearly 280 world-class giant fields thus far discovered, plus the supergiants, account for about 80 percent of the world's known recoverable oil. There are, in addition, approximately 1,000 known large oil fields that initially contained between 50,000,000 and 500,000,000 barrels. These fields account for some 14 to 16 percent of the world's known oil. Less than 5 percent of the known fields originally contained roughly 95 percent of the world's known oil.

Sedimentary basins. Giant petroleum fields and significant petroleum-producing sedimentary basins are closely associated. In some basins, huge amounts of petroleum apparently have been generated because perhaps only about 10 percent of the generated petroleum is trapped and preserved. The Arabian-Iranian sedimentary basin is predominant because it contains more than 20 supergiant fields. No other basin has more than one such field. In 20 of the 26 most significant oil-containing basins, the 10 largest fields originally contained more than 50 percent of the known recoverable oil. Known world oil reserves are concentrated in a relatively small number of giant fields in a few sedimentary basins.

Worldwide, approximately 600 sedimentary basins are known to exist. About 160 of these have yielded oil, but only 26 are significant producers and 7 of these account for more than 65 percent of total known oil. Exploration has occurred in another 240 basins, but discoveries of commercial significance have not been made.

Geologic study and exploration. Current geologic understanding can usually distinguish between geologically favourable and unfavourable conditions for oil accumulation early in the exploration cycle. Thus, only a relatively few exploratory wells may be necessary to indicate whether a region is likely to contain significant amounts of oil. Modern petroleum exploration is an efficient process. If giant fields exist, it is likely that most of the oil in a region will be found by the first 50 to 250 exploratory wells. This number may be exceeded if there is a much greater than normal amount of major prospects or if exploration drilling patterns are dictated by either political or unusual technological considerations. Thus, while undiscovered commercial oil fields may exist in some of the 240 explored but seemingly barren basins, it is unlikely that they will be of major importance since the largest are normally found early in the exploration process.

The remaining 200 basins have had little or no exploration, but they have had sufficient geologic study to indicate their dimensions, amount and type of sediments, and general structural character. Most of the underexplored (or frontier) basins are located in difficult environments, such as polar regions or submerged continental margins. The larger sedimentary basins—those containing more than 833,000 cubic kilometres (200,000 cubic miles) of sediments—account for some 70 percent of known world petroleum. Future exploration will have to involve the smaller basins as well as the more expensive and difficult frontier basins.

Status of the world oil supply. The first 200,000,000,000 barrels of world oil were produced in 109 years from 1859 to 1968. Since that time world oil production rates have stabilized at a rate of about 22,000,000,000 barrels a year.

Table 1 shows the broad distribution of the world oil supply. Reserves are identified quantities of "in-place" petroleum that are considered recoverable under current economic and technological conditions. Estimated by petroleum engineers and geologists using drilling and production data along with other subsurface information, these figures are revised to include projected field growth as development progresses. Petroleum reserves are reported by oil companies and by some governments, and such data are compiled by the U.S. Department of Energy and the U.S. Geological Survey, as well as by oil industry trade journals. Undiscovered petroleum resources of the world have been estimated by the U.S. Geological Survey by the extrapolation of known production and reserve data into untested sediments of similar geology. A most likely consensus estimate was established, as was a range with upper and lower yield limits at 5 and 95 percent probabilities. The range for undiscovered oil resources assessed for the whole world is 275,000,000,000 to 1,469,000,000,000 barrels.

Table 2: Leading Oil Countries

country	cumulative production*	reserves*	undiscovered resources*	total oil endowment*
Saudi Arabia	71.5	261.2	41	373.7
United States	165.8	50.7	49	265.5
Russia	92.6	100.0	68	260.6
Iraq	22.8	100.0	45	167.8
Iran	42.9	93.0	22	157.9
Venezuela	47.3	83.3	17	147.6
Kuwait	27.6	97.5	3	128.1
United Arab Emirates	15.1	98.2	7	120.3
Mexico	20.5	50.4	37	107.9
China	18.8	24.0	48	90.8
Canada	16.1	5.1	33	54.2
Libya	19.0	22.8	8	49.8
Kazakhstan	3.2	17.3	26	46.5
Nigeria	15.5	17.9	9	42.4
Indonesia	15.2	5.8	10	31.0
Norway	6.3	11.3	13	30.6
United Kingdom	12.3	4.6	11	27.9
Algeria	9.1	9.2	2	20.3
Totals	621.6	1,052.3	449	2,122.9

*In billion barrels; figures adapted from *Oil & Gas Journal* and U.S. Geological Survey.

Duration of oil supply

Table 1 indicates that the most likely total world oil endowment is about 2,390,000,000,000 barrels. Of this amount, 77 percent has already been discovered and 30 percent has already been produced and consumed. If this estimate proves to be reasonably accurate, current relatively stabilized world oil-production volumes could be sustained to about the middle of the 21st century, at which time a shortage of conventional oil resources would force a production decline.

The Middle East is thought to have had an estimated 41 percent of the world's total oil endowment. North America is a distant second but has already produced almost half of

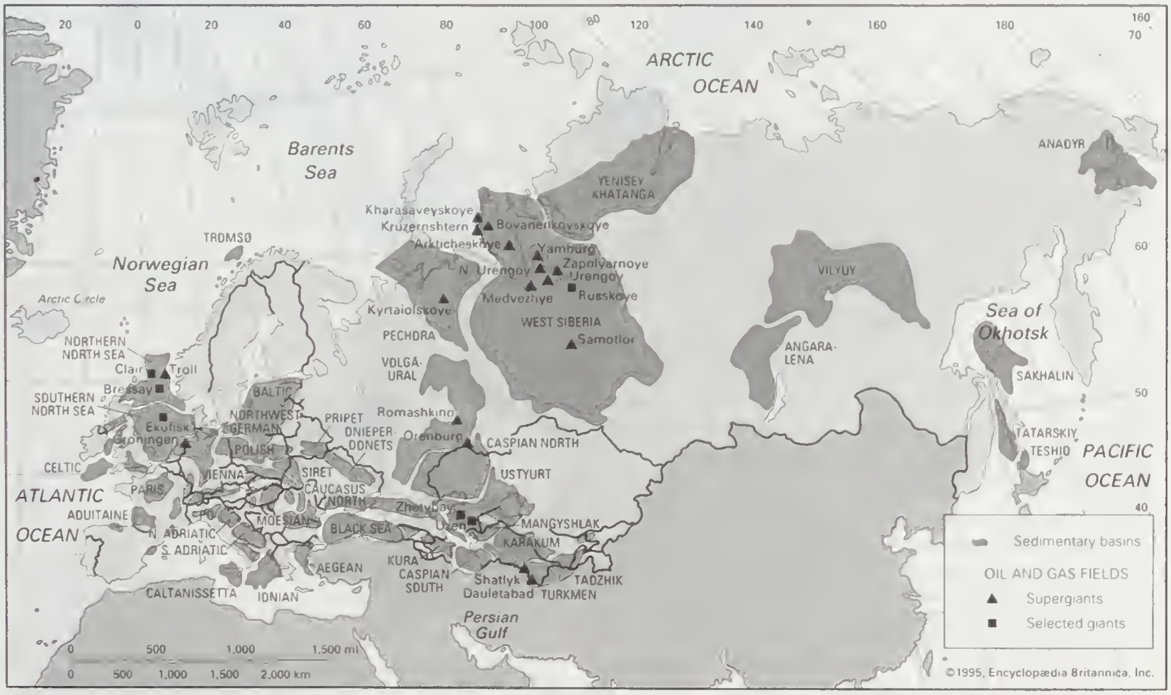


Figure 4: Sedimentary basins and major oil and gas fields of Europe, Russia, Transcaucasia, and Central Asia.

its total oil. Eastern Europe, because of the large deposits in Russia, is well endowed with oil. Western Europe is not, with most of its oil under the North Sea. Likewise, Africa, Asia, and South America are thought to have only relatively moderate amounts of oil. It is interesting to note that a large undiscovered oil resource is believed to exist in North America, which has many frontier basins. Both the Middle East and eastern Europe, however, are also thought to contain significant oil prospects.

Major producing countries. Table 2 lists the 18 countries believed to have had an original oil endowment exceeding 20,000,000,000 barrels. It also serves to show the concentration of world oil. These 18 countries have accounted for 86 percent of the world's oil production. They hold 94 percent of its reserves. Significantly, they are projected to have 82 percent of the world's remaining undiscovered oil resources. As can be seen, regions geologically favourable to the generation and deposition of oil are fairly rare. The 18 countries listed are estimated to have contained 89 percent of the world's original oil endowment.

Saudi Arabia. Saudi Arabia, shown in Figure 3, is thought to have had the largest original oil endowment of any country. The discovery that transformed Saudi Arabia into a leading oil country was the Al-Ghawār field. Discovered in 1948, this field has proved to be the world's largest, containing 82,000,000,000 barrels. Another important discovery was the Saffāniyah offshore field in the Persian Gulf. It is the third largest oil field in the world and the largest offshore. Saudi Arabia has eight other supergiant oil fields. Thus, it has the largest oil reserve in the world, not to mention significant potential for additional discoveries.

Russia. Russia is thought to possess the best potential for new discoveries. Also, it has significant reserves. Russian oil is derived from many sedimentary basins within the vast country, while Saudi Arabian fields, as well as many other Middle Eastern fields, are located in the great Arabian-Iranian basin (Figures 3 and 4). Russia has two supergiant oil fields, Samotlor and Romashkino. Production from these fields is on the decline, bringing total Russian oil output down with them. The best prospects for new Russian discoveries appear to exist in the difficult and expensive frontier areas.

United States, Mexico, and Canada. North America also has many sedimentary basins; they are shown in Figure 5. Basins in the United States have been intensively explored and their oil resources developed. More than 33,000 oil fields have been found, but only two are supergiants (Prudhoe Bay in the North Slope region of Alaska, shown in Figure 6, and East Texas). Cumulatively, the United States has produced more oil than any other country but is still considered to have a significant remaining undiscovered oil resource. Prudhoe Bay, which accounted for approximately 17 percent of U.S. oil production dur- Prudhoe Bay

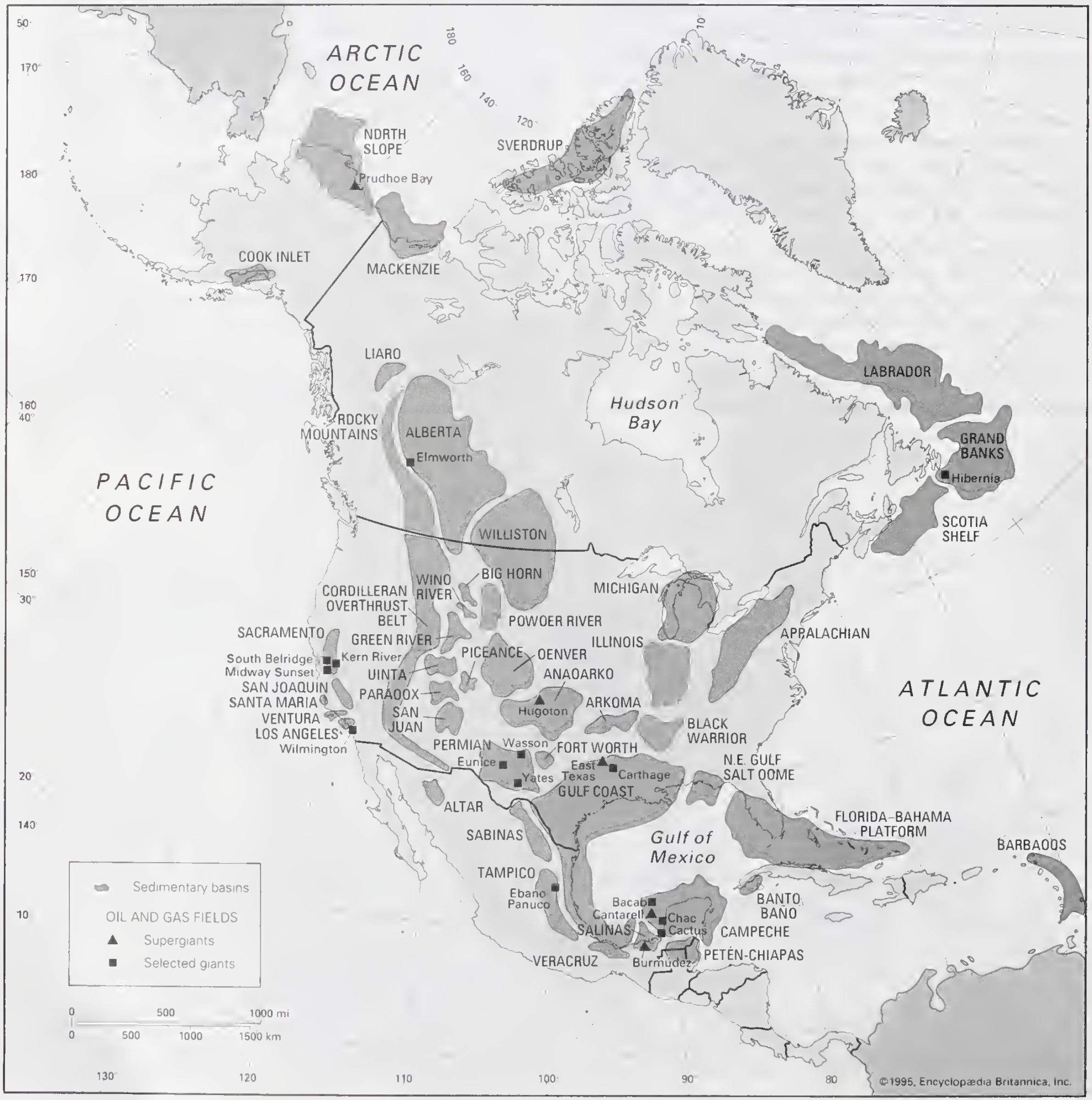


Figure 5: Sedimentary basins and major oil and gas fields of North America.

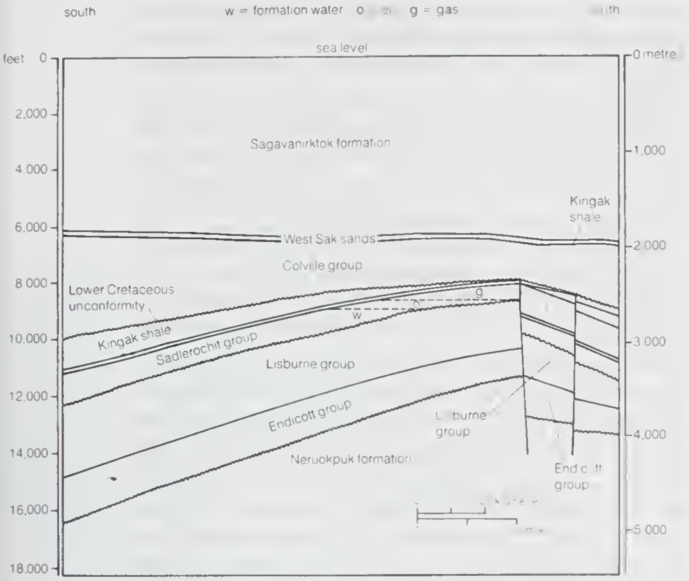


Figure 6: North-to-south cross section of the Prudhoe Bay complex.

From H C Jamison L D Brockett, and R A McIntosh, 'Prudhoe Bay—A 10-Year Perspective, in Michel T Halbouty ed , *Giant Oil and Gas Fields of the Decade 1968–1978*, The American Association of Petroleum Geologists, Tulsa Okla Memoir 30, 1980

ing the mid-1980s, is in decline. This situation, coupled with declining oil production in the conterminous United States, has contributed to a significant drop in domestic oil output. Mexico has produced only about one-fifth of its estimated total oil endowment. With two supergiant fields (Cantarell offshore of Campeche state and Bermudez in Tabasco state) and with substantial remaining reserves and resources, it will be able to sustain current production levels well into the 21st century. Conversely, Canada, with considerably smaller oil reserves and most of its undiscovered resource potential in remote regions, is unlikely to be able to sustain current production levels beyond the 1990s. Canada's largest oil field is Hibernia, discovered off Newfoundland in 1979. This giant field has yet to be developed.

Iraq, Kuwait, and Iran. The Middle Eastern countries of Iraq, Kuwait, and Iran are each estimated to have had an original oil endowment in excess of 100,000,000,000 barrels. These countries have a number of supergiant fields, all of which are located in the Arabian-Iranian basin, including Kuwait's Al-Burqān field (Figure 3). Al-Burqān is the world's second largest oil field, having originally contained 75,000,000,000 barrels of recoverable oil. Iraq possesses a significant potential for additional oil discoveries.

United Kingdom. The United Kingdom is an important North Sea exporter; however, as its undiscovered resource potential appears somewhat limited, it may require more of its oil output for internal use in the future.

Undiscovered resources. With an estimated 77 percent of the world's total recoverable oil endowment having already been discovered, the remaining 23 percent, mostly located in smaller fields or in more difficult environments, is expected to become ever more expensive to find and to recover. More than 11,000 man-years were required to construct the largest of the North Sea gravity platforms, making capital costs per daily oil production as much as 40 times the costs in the Middle East. A guyed tower constructed in more than 300 metres of water in the Gulf of Mexico has been estimated to produce oil at about 65 times the production cost in the Middle East. As oil exploitation moves into deeper waters or under Arctic ice, the cost will further escalate and will be reflected in the world economy. (J.P.Ri./G.I.A.)

Natural gas

Natural gas is often found dissolved in oil at the high pressures existing in a reservoir, and it also can be present as a gas cap above the oil. Such natural gas is known as associated gas. There are also reservoirs that contain gas and no oil. This gas is termed nonassociated gas.

Associated and non-associated gas

HISTORY OF USE

Discovery and early application. The first discoveries of natural gas seeps were made in Iran between 6000 and 2000 BC. Many early writers described the natural petroleum seeps in the Middle East, especially in the Baku region of what is now Azerbaijan. The gas seeps, probably first ignited by lightning, provided the fuel for the "eternal fires" of the fire-worshiping religion of the ancient Persians. The use of natural gas was mentioned in China about 900 BC. It was in China in 211 BC that the first known well was drilled for natural gas to reported depths of 150 metres (500 feet). The Chinese drilled their wells with bamboo poles and primitive percussion bits for the express purpose of searching for gas in Late Triassic limestones (more than 208,000,000 years old) in an anticline west of modern Chungking. The gas was burned to dry the rock salt found interbedded in the limestone. Eventually wells were drilled to depths approaching 1,000 metres, and more than 1,100 wells had been drilled into the anticline by 1900.

Natural gas was unknown in Europe until its discovery in England in 1659, and even then it did not come into wide use. Instead, gas obtained from carbonized coal (known as town gas) became the primary fuel for illuminating streets and houses throughout much of Europe from 1790 on. In North America the first commercial application of a petroleum product was the utilization of natural gas from a shallow well in Fredonia, N.Y., in 1821. The gas was distributed through a small-bore lead pipe to consumers for lighting and cooking.

Improvements in gas pipelines. Throughout the 19th century the use of natural gas remained localized because there was no way to transport large quantities of gas over long distances. Natural gas remained on the sidelines of industrial development, which was based primarily on coal and oil. An important breakthrough in gas-transportation technology occurred in 1890 with the invention of leakproof pipeline coupling. Nonetheless, materials and construction techniques remained so cumbersome that gas could not be used more than 160 kilometres (100 miles) from a source of supply. Thus, associated gas was mostly flared (*i.e.*, burned at the wellhead), and nonassociated gas was left in the ground, while town gas was manufactured for use in the cities.

Long-distance transmission of natural gas

Long-distance gas transmission became practical during the late 1920s because of further advances in pipeline technology. From 1927 to 1931 more than 10 major transmission systems were constructed in the United States. Each of these systems was equipped with pipes having diameters of approximately 51 centimetres (20 inches) and extended more than 320 kilometres. Following World War II, a large number of even longer pipelines of increasing diameter were constructed. The fabrication of pipes having a diameter of up to 142 centimetres became possible. Since the early 1970s the longest gas pipelines have had their origin in Russia. For example, the 5,470-kilometre-long Northern Lights pipeline crosses the Ural Mountains and some 700 rivers and streams, linking eastern Europe with the West Siberian gas fields on the Arctic Circle. As a result, gas from the Urengoy field, the world's largest, is transported to eastern Europe and then on to western Europe for consumption. Another gas pipeline, shorter but also of great engineering difficulty, is the 51-centimetre line that extends from Algeria across the Mediterranean Sea to Sicily. The sea is more than 600 metres deep along some parts of the route.

Natural gas as a premium fuel. As recently as 1960, associated gas was a nuisance by-product of oil production in many areas of the world. The gas was separated from the crude oil stream and eliminated as cheaply as possible, often by flaring. Only since the crude oil shortages of the late 1960s and early 1970s has natural gas become an important world energy source.

Even in the United States the home-heating market for natural gas was limited until the 1930s when town gas began to be replaced by abundant and cheaper supplies of natural gas, which contained twice the heating value of its synthetic predecessor. Also, when natural gas burns completely, carbon dioxide and water are normally formed. The combustion of gas is relatively free of soot, carbon monoxide, and the nitrogen oxides associated with the burning of other fossil fuels. In addition, sulfur dioxide emissions, another major air pollutant, are almost nonexistent. As a consequence, natural gas is often a preferred fuel for environmental reasons.

PROPERTIES

Chemical composition. *Hydrocarbon content.* Natural gas is a hydrocarbon mixture consisting primarily of methane and ethane, both of which are gaseous under atmospheric conditions. The mixture also may contain other hydrocarbons, such as propane, butane, pentane, and hexane. In natural gas reservoirs even the heavier hydrocarbons occur for the most part in gaseous form because of the higher pressures. They liquefy at the surface (at atmospheric pressure) and are referred to as natural gas liquids, gas condensate, natural gasoline, or liquefied petroleum gas. They may separate in some reservoirs through retrograde condensation or may be separated at the surface either in field separators or in gas processing plants by means of condensation, absorption, adsorption, or other modification. The average production of natural gas liquids in the United States is nearly 38 barrels per 1,000,000 cubic feet of produced gas.

Natural gas liquids

Nonhydrocarbon content. Other gases that commonly occur in association with the hydrocarbon gases are nitrogen, carbon dioxide, hydrogen, hydrogen sulfide, and such noble gases as helium and argon. Because natural gas and formation water occur together in the reservoir, gas recovered from a well contains water vapour, which is partially condensed during transmission to the processing plant.

Physical properties. The physical properties of natural gas include colour, odour, and flammability. The principal ingredient of gas is methane, which is colourless, odourless, and highly flammable. However, some of the associated gases in natural gas, especially hydrogen sulfide, have a distinct and penetrating odour, and a few parts per million is sufficient to impart a decided odour to natural gas.

Measurement systems. The amounts of gas accumulated in a reservoir, as well as produced from wells, are calculated in cubic metres at a pressure of 750 millimetres of mercury and a temperature of 15° C (or in cubic feet at an absolute pressure of 14.73 pounds per square inch and a temperature of 60° F). Since gas is compressed at high reservoir pressures, it expands upon reaching the surface and thus occupies more space. As its quantity is calculated in reference to standard conditions of temperature and pressure, however, the expansion does not constitute an increase in the amount of gas produced.

ORIGIN

Organic formation process. Natural gas is more ubiquitous than oil. It is derived from both land plants and aquatic organic matter and is generated above, throughout, and below the oil window. Thus, all source rocks have the potential for gas generation. Many of the source rocks for significant gas deposits appear to be associated with the worldwide occurrence of Carboniferous coal (roughly 286,000,000 to 360,000,000 years in age).

The biological stage. During the immature, or biological, stage of petroleum formation, biogenic methane (often called marsh gas) is produced as a result of the decomposition of organic material by the action of anaerobic microbes. These microorganisms cannot tolerate even traces of oxygen and are also inhibited by high concentrations of dissolved sulfate. Consequently, biogenic gas generation is confined to certain environments that include poorly drained swamps and bays, some lake bottoms, and marine environments beneath the zone of active sulfate reduction. Gas of predominantly biogenic origin is thought to constitute more than 20 percent of the world's gas reserves.

Biogenic methane

The mature stage of petroleum generation, which occurs at depths of about 760 to 4,880 metres, includes the full range of hydrocarbons that are produced within the oil window. Often significant amounts of thermal methane gas are generated along with the oil. Below 2,900 metres, primarily wet gas (gas containing liquid hydrocarbons) is formed.

The thermal stage. In the postmature stage, below about 4,880 metres, oil is no longer stable, and the main hydrocarbon product is thermal methane gas. The thermal gas is the product of the cracking of the existing liquid hydrocarbons. Those hydrocarbons with a larger chemical structure than that of methane are destroyed much more rapidly than they are formed. Thus, in the sedimentary basins of the world, comparatively little oil is found below 4,880 metres. The deep basins with thick sequences of sedimentary rocks, however, have the potential for deep gas production.

Inorganic formation. Some methane may have been produced by inorganic processes. The original source of the Earth's carbon was the cosmic debris from which the planet formed. If meteorites are representative of this debris, the carbon could have been supplied in comparatively high concentrations as hydrocarbons, such as are found in the carbonaceous chondrite type of meteorites. Continuous outgassing of these hydrocarbons may be taking place from within the Earth, and some may have accumulated as abiogenic gas deposits without having passed through an organic phase. In the event of widespread outgassing, however, it is likely that abiogenic gas would be too diffuse to be of commercial interest. Significant accumulations of inorganic methane have yet to be found.

The helium and some of the argon found in natural gas are products of natural radioactive disintegration. Helium derives from radioisotopes of thorium and the uranium family, and argon derives from potassium. It is probably coincidental that helium and argon sometimes occur with natural gas; in all likelihood, the unrelated gases simply became caught in the same trap. (J.P.Ri./Ed.)

The geologic environment. Like oil, natural gas migrates and accumulates in traps (Figure 1). Oil accumulations contain more recoverable energy than gas accumulations of similar size, even though the recovery of gas is a more efficient process than the recovery of oil. This is due to the differences in the physical and chemical properties of gas and oil. Gas displays initial low concentration and high dispersibility, making adequate cap rocks very important.

Natural gas can be the primary target of either deep or shallow drilling because large gas accumulations form above the oil window as a result of biogenic processes and thermal gas occurs throughout and below the oil window. In most sedimentary basins the vertical potential (and sediment volume) available for gas generation exceeds that of oil. About a quarter of the known major gas fields are related to a shallow biogenic origin, but most major gas fields are located at intermediate or deeper levels where higher temperatures and older reservoirs (often carbonates sealed by evaporites) exist.

Vertical extent of gas formations

Conventional gas reservoirs. Gas reservoirs differ greatly, with different physical variations affecting reservoir performance and recovery. In a natural gas (single-phase) reservoir it should be possible to recover nearly all of the in-place gas by dropping the pressure sufficiently. If the pressure is effectively maintained by the encroachment of water in the sedimentary rock formation, however, some of the gas will be lost to production by being trapped by capillarity behind the advancing water front. Therefore, in practice, only about 80 percent of the in-place gas can be recovered. On the other hand, if the pressure declines, there is an economic limit at which the cost of compression exceeds the value of the recovered gas. Depending on formation permeability, actual gas recovery can be as high as 75 to 80 percent of the original in-place gas in the reservoir. Associated gas is produced along with the oil and separated at the surface.

Unconventional gas reservoirs. Substantial amounts of gas have accumulated in geologic environments that differ from conventional petroleum traps. This gas is termed unconventional gas and occurs in "tight" (*i.e.,* relatively impermeable) sandstones, in joints and fractures or absorbed into the matrix of shales (often of the Devonian Period [about 360,000,000 to 408,000,000 years old]), dissolved or entrained in hot geopressured formation waters, and in coal seams. Unconventional gas sources are much more expensive to exploit and have to be produced at much slower rates than conventional gas fields. Moreover, recoveries are low. In all likelihood, unconventional gas will continue to complement conventional gas production but will not supplant it.

WORLD DISTRIBUTION

Location of major gas fields. The largest natural gas fields are the supergiants, which contain more than 850,000,000,000 cubic metres of gas, and the world-class giants, which have reserves of roughly 85,000,000,000 to 850,000,000,000 cubic metres. Supergiants and world-class giants represent less than 1 percent of the world's total known gas fields, but they originally contained, along with associated gas in giant oil fields (see above), approximately 80 percent of the world's reserves and produced gas.

Urengoy

Russia. Some of the world's largest gas fields occur in Russia in a region of West Siberia east of the Gulf of Ob on the Arctic Circle. The world's largest gas field is Urengoy, which was discovered in 1966. Its initial reserves have been estimated at 8,087,000,000,000 cubic metres. Nearly 6,230,000,000,000 cubic metres of this gas are in the shallowest reservoir, 1,100 to 1,250 metres deep, which is Upper Cretaceous in age (from about 66,400,000 to 97,500,000 years old). In all, Urengoy has 15 separate reservoirs, some in Lower Cretaceous rocks (those that are approximately 97,500,000 to 144,000,000 years old). The deepest is a gas condensate zone in Upper Jurassic strata (from about 144,000,000 to 163,000,000 years old). Urengoy began production in 1978. Its maximum output is expected to be as much as 250,000,000,000 cubic metres of gas per year, which would considerably exceed the production from any other gas field in the world.

Yamburg, Russia's second largest gas field, was discovered north of the Arctic Circle and north of Urengoy. Its original reserves were estimated at 4,700,000,000,000 cubic metres of gas, mostly from Upper Cretaceous reservoir rocks at depths of 1,000 to 1,210 metres. Development of Yamburg began in the early 1980s. Bovanenkovskoye, discovered in 1971 on the Yamal Peninsula in northwestern Siberia, is Russia's third largest field. It has reserves estimated at 4,102,000,000,000 cubic metres in Lower Cretaceous reservoir rock at depths of 1,190 to 1,475 metres. Bovanenkovskoye has not yet been developed. Some of the gas in these huge, shallow fields may be of biogenic origin and capped by permafrost.

Orenburg, discovered in the Volga-Urals region in 1967, is the largest Russian gas field outside of West Siberia. It had initial reserves of 1,778,300,000,000 cubic metres of gas and is now under production.

Europe. The largest natural gas field in Europe is Groningen, with original recoverable reserves of about 2,270,000,000,000 cubic metres. It was discovered in 1959 on the Dutch coast. The discovery well was drilled through evaporites of Permian age (about 245,000,000 to 286,000,000 years old) into a thick basal Permian sandstone that was gas-productive. Subsequent drilling outlined a broad anticline about 24 kilometres wide by 40 kilometres long, which has a continuous basal Permian sandstone reservoir capped by evaporites. The reservoir contains natural gas at depths between 2,440 and 3,050 metres. It overlies the truncated and strongly faulted coal-bearing Pennsylvanian sequence (the Pennsylvanian Period extended from about 286,000,000 to 320,000,000 years ago), which is considered to be the main source of the gas.

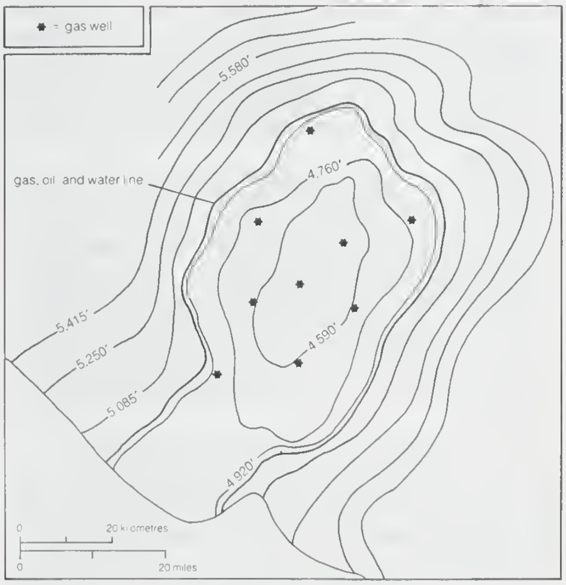


Figure 7: Reservoir of Hassi R'Mel gas field, Algeria.
From Philippe R. Magloire, "Triassic Gas Field of Hassi er R'Mel, Algeria," in Michel T. Halbouty ed., *Geology of Giant Petroleum Fields*, p. 496, The American Association of Petroleum Geologists, Tulsa Okla. Memoir 14, 1970

Hugoton

North America. In the United States, Hugoton, discovered in 1927 in Kansas and found to extend through the Oklahoma and Texas panhandles, is a gas field with an estimated ultimate recovery of 1,986,000,000,000 cubic metres. More than 7,000 wells have been drilled in this extensive field, which produces from a series of Permian limestones and dolomites. The gas accumulations are stratigraphically controlled by variations in lithology. The productive area extends along a 400-kilometre trend. Canada has a significant estimated endowment of natural gas, of which only about 17 percent has been produced. Its undiscovered resource potential is almost equal to that of the United States. The largest gas field is Elmworth. Discovered in Alberta in 1976, Elmworth contained some 560,000,000,000 cubic metres of gas in a Cretaceous sandstone reservoir. Mexico's largest gas accumulation is associated with the supergiant Bermudez oil field. Located in 1958 in the Chiapas-Tabasco region, Bermudez originally contained 490,000,000,000 cubic metres of associated gas in a Cretaceous dolomite reservoir. Although Mexico's estimated gas endowment is less than half that of Canada, natural gas is underutilized in Mexico, and only 11 percent of that country's estimated total recoverable gas has been produced.

North Africa. In North Africa the central basin of Algeria is the location of the Hassi R'Mel gas and condensate field (Figure 7). Discovered in 1956 in a large anti-

Table 3: The Recoverable Natural Gas Resources of the World*

region	cumulative production	reserves	undiscovered resources	total gas endowment
United States	22.4	4.6	11.2	38.2
Canada	2.6	2.7	10.3	15.6
Mexico	0.8	2.0	4.4	7.2
South America	1.8	5.5	5.9	13.2
Western Europe	4.1	5.4	5.8	15.3
Russia and Ukraine	8.6	47.0	45.0	100.6
Transcaucasia and Central Asia	2.9	10.7	6.6	20.2
Middle East	2.1	44.3	31.5	77.9
Africa (including North Africa)	1.1	9.6	12.4	23.1
China	0.5	1.7	7.3	9.5
Oceania and Asia (excluding China)	2.0	8.3	13.0	23.3
Total world	48.9	141.8	153.4	344.1

*In trillion cubic metres; figures adapted from *Oil & Gas Journal* and U.S. Geological Survey.

cline, the field is estimated to have originally contained about 2,520,000,000,000 cubic metres of recoverable gas in reservoirs of permeable Triassic sandstone (from about 208,000,000 to 245,000,000 years old) capped by salt beds. Hassi R'Mel is under development and is reported to have the capacity to produce 59,000,000,000 cubic metres of gas per year.

Middle East. There is an enormous gas potential in the Middle East associated with the major structures in the Arabian-Iranian basin. The Permian Khuff formation underlies most of the region and is an important gas-bearing horizon. It forms the reservoir of the supergiant North Field of offshore Qatar and also of other smaller nonassociated gas fields in the region. There also is great potential for nonassociated gas accumulations in Lower Cretaceous (as well as in the Permian) strata should the demand for Persian Gulf gas rise, either for domestic use or for export.

Asia. The largest gas field in Asia is Arun, which was discovered in 1971 in the North Sumatra basin of Indonesia. The gas reservoir is a reef limestone of middle Miocene age (from about 11,000,000 to 16,000,000 years old). Original reserves have been estimated at about 383,000,000,000 cubic metres. The gas is liquefied for export.

Status of world reserves. When the generation and migration of gas is considered, the extensive vertical gas-generation zone includes shallow biogenic gas, the intermediate dissolved gas of the oil window, and deeper thermal gas. This large vertical habitat for gas and the additional availability of source material indicate that considerable gas may have been formed and still remains undiscovered. Table 3, derived from an assessment of the U.S. Geological Survey and other estimates in the technical literature, shows the broad distribution of world natural gas. It is estimated that 45 percent of the world's recoverable gas remains undiscovered and that, on the basis of energy content, the world's ultimate recoverable resources of natural gas will approach those of oil. Because the utilization of gas in large volumes lags behind the use of oil, the world's stock of gas is expected to last longer than that of oil. However, if the consumption of gas approaches that of oil on an equivalent basis, it, too, will be short-lived as a major energy resource.

Duration of gas supply

About 14 percent of the world's estimated total gas endowment has been consumed or flared. The flaring of associated gas has long been a practice connected with oil production. As recently as 1980, approximately 10 percent of world annual gas production was lost at the wellhead by this procedure. Historically, Middle Eastern and African oil-producing countries have flared the most gas. Much of the gas yielded is reinjected, but what cannot be reinjected has often been flared because the remote location of many oil wells makes the recovery of gas expensive. As the value of gas has appreciated, however, conservation efforts have increased and gas flaring has been reduced.

Table 3 shows that the estimated total world endowment of natural gas is more than 344,000,000,000,000 cubic metres. About one-third of this gas was originally located in the Soviet Union, which, prior to its dissolution in 1991, had surpassed the United States to become the world's leading producer of natural gas. Together, the Soviet Union and the Middle East originally accounted for more than half of the world's natural gas endowment. The United States also possessed a significant endowment of natural gas, but it has already consumed more than half of its resources. U.S. gas production has been projected to fall by as much as 10 percent by the end of the 20th century because of the declining resource base.

The total gas endowments of Latin America, western Europe, Africa, and Asia and the Pacific region, while significant, are thought to be considerably smaller than those of North America, the former Soviet Union, and the Middle East. However, past gas production in these regions has been somewhat limited; therefore much of the original gas is still available for use.

Russia had the world's largest original gas endowment—more than 98,000,000,000,000 cubic metres. The United States and Iran both had original gas endowments of more than 33,000,000,000,000 cubic metres. The gas endowments of the following countries were in excess of 2,800,000,000,000 cubic metres in descending order: Saudi Arabia, Canada, China, Turkmenistan, Norway, Mexico, the United Arab Emirates, Nigeria, Qatar, Kazakhstan, Venezuela, Indonesia, Kuwait, Australia, Algeria, Uzbekistan, Malaysia, The Netherlands, and Ukraine. These countries originally possessed more than 90 percent of the world's total recoverable natural gas.

Heavy oils and tar sands

Specific gravity of heavy oils

Crude oils below 20° API gravity are usually considered to be heavy. The lighter conventional crudes are often waterflooded to enhance recovery. The injection of water into the reservoir helps to maintain reservoir pressure and displace the oil toward the production wells. In general, waterflooding is most effective with light crude oils of API gravity 25° and higher and becomes progressively less effective with oils below 25° API. With crudes of 20° and lower, waterfloods are essentially ineffective and thermal recovery becomes necessary. Very few thermal projects are successful in recovering oil of less than 10° API gravity. Heavy crude oils have enough mobility that, given time, they will be producible through a well bore in response to thermal recovery methods. Tar sands contain immobile bitumen that will not flow into a well bore even under thermal stimulation. The recovery of these resources requires mining.

HISTORY OF USE

Discovery. In ancient times the Elamites, Chaldeans, Akkadians, and Sumerians mined shallow deposits of asphalt, or bitumen, for their own use. Mesopotamian bitumen was exported to Egypt where it was employed for various purposes, including the preservation of mummies. The Dead Sea was known as Lake Asphaltites (from which the term asphalt was derived) because of the lumps of semisolid petroleum that were washed up on its shores from underwater seeps.

Bitumen had many other uses in the ancient world. It was mixed with sand and fibrous materials for use in the construction of watercourses and levees and as mortar for bricks. It was widely used for caulking ships and in road building. Bitumen also was employed for bonding tools, weapons, and mosaics and in inlaid work and jewel setting. In various areas it was used in paints and for waterproofing baskets and mats. Artistic and religious objects were carved from bitumen-impregnated sands, and the mining of rock asphalt was an important industry.

Centuries later, during the age of exploration, Sir Walter Raleigh found the famous "Pitch Lake" deposits in Trinidad. The Dutch made similar discoveries in Java and Sumatra.

Potential as a crude oil source. In response to thermal recovery methods, world production of heavy oil exceeds 1,000,000 barrels per day, or roughly 2 percent of the total world oil output. In the United States approximately 6 percent of total oil production is derived from heavy oil fields. The production of synthetic oil from the bitumen in tar sands is limited to Alberta, Can., and amounts to about 250,000 barrels per day.

Yet, the development of heavy oil and bitumen reserves is increasing around the world. The increasing volume of cheaper heavy oil in the supply mix has provided an incentive for refiners to upgrade their equipment to process the poorer-quality heavier crudes. The upgrading investments have helped to maintain a demand for heavy oil in spite of the declining price of conventional crudes since the early 1980s. As the demand for heavy oil and syncrude from tar sands remains strong, heavy-hydrocarbon development projects are being initiated in several parts of the world. In addition, unsuccessful attempts to find new giant conventional oil fields in recent years has caused some producers to turn to the marginally economic heavy hydrocarbons to replace depleted reserves.

COMPOSITION AND ORIGIN

Chemical composition. Geochemical analyses indicate that the heavy hydrocarbons are composed primarily of as-

Table 4: The Recoverable Heavy Oil Resources of the World*

region	cumulative production	reserves	undiscovered resources	total heavy oil
United States	10.8	17.4	23.6	51.8
Canada	0.4	0.7	4.9	6.0
Mexico	1.2	4.2	1.6	7.0
Venezuela	13.2	136.7	13.0	162.9
Remaining Latin America	0.2	1.6	3.2	5.0
Western Europe	0.8	8.0	0.2	9.0
Eastern Europe	0.1	0.3	0	0.4
Russia	4.7	6.2	18.5	29.4
Transcaucasia and Central Asia	0.5	0.7	2.1	3.3
Middle East	32.5	114.4	22.1	169.0
Africa (including North Africa)	0.4	3.6	0.6	4.6
China	0.3	9.2	0	9.5
Oceania and Asia (excluding China)	2.3	3.7	3.6	9.6
Total world	67.4	306.7	93.4	467.5

*In billion barrels; figures adapted from U.S. Geological Survey, American Association of Petroleum Geologists, and *Oil & Gas Journal.*

phaltenes, resins, and metals (most commonly vanadium and nickel). The nature of individual heavy oil deposits varies widely as they are rarely chemically homogeneous. Bitumen distribution in a deposit also varies, depending on the permeability and porosity of the reservoir rock.

Degradation of conventional oils

Formation. Nearly all the deposits of heavy hydrocarbons are degraded remnants of accumulations of conventional oils. Degradation begins when oil migrates toward the surface and encounters descending meteoric water (rainwater or any other water of atmospheric origin) containing oxygen and bacteria at temperatures below 93° C. A tarlike material is formed at the oil-water contact, and it eventually invades the entire oil accumulation. A process known as "water washing" removes the more water-soluble light hydrocarbons, particularly the aromatics. Biodegradation preferentially removes the normal paraffins. Heavy hydrocarbon accumulations may represent as little as 10 percent of the original conventional oil. They contain asphaltenes, resins, sulfur, and such metals as vanadium and nickel, which results in an increase in density. These apparently are the residues of a natural concentrating process and were not contributed by other sources. Thus, the deposits were emplaced as medium-gravity crudes, which later became immobilized by degradation in the reservoir. Some of the heavy oils, however, appear thermally immature and therefore may be unaltered.

The geologic environment. Almost all the heavy hydrocarbon deposits have been found in formations of Cretaceous and Tertiary age (about 1,600,000 to 144,000,000 years old). The exceptions include some deposits in Alberta, Can., and in Russia. In Alberta bituminous Paleozoic carbonates unconformably underlie Mesozoic rocks (the Paleozoic Era began about 540,000,000 years ago and lasted until the beginning of the Mesozoic Era, roughly 245,000,000 years ago). In Russia most of the heavy hydrocarbons occur in strata dating back to the Paleozoic Era and earlier (*i.e.*, the late Precambrian, which ended about 540,000,000 years ago). Some heavy hydrocarbons are found in Tertiary rocks in Central Asia.

The most prolific heavy hydrocarbon reservoir sediments are sandstones that were originally deposited in fluvial and deltaic, nearshore environments. The exceptions are the bituminous carbonate rocks of Alberta, Russia, and Central Asia. Smaller deposits of asphaltic carbonate rocks are common, notably in the Middle East and in Italy. Many heavy oil reservoirs have been found offshore beneath the continental shelves of Africa and North and South America. In addition, heavy hydrocarbons have been discovered beneath the Caspian, Mediterranean, Adriatic, Red, Black, North, Beaufort, and Caribbean seas, as well as beneath the Persian Gulf and the Gulf of Mexico.

WORLD DISTRIBUTION

Commercially viable deposits. *Heavy oil.* California accounts for nearly all of the thermally recovered heavy oil in the United States. The largest of the California heavy oil fields is Midway-Sunset, with an ultimate recovery estimated at slightly more than 2,000,000,000 barrels. Almost as large is the Kern River field, projected to ultimately produce slightly less than 2,000,000,000 barrels. Two other California fields, Wilmington and South Belridge, are estimated to have originally contained more than 1,000,000,000 barrels of recoverable heavy oil. There are five additional heavy oil fields estimated to have initially contained more than 500,000,000 barrels.

Four of the world's largest oil fields, the supergiants Al-Burqān in Kuwait, Kirkuk in Iraq, Abū Sa'fah in Saudi Arabia, and the Bolivar Coastal field in Venezuela, contain and have produced very large amounts of heavy oil in addition to conventional oils. Giant fields producing heavy oil include Zubair in Iraq; Duri in Indonesia; Gudao and Karamai in China; Seria in Brunei; Bacab, Chac, and Ebano-Panuco in Mexico; Belayim Land in Egypt; Maydan Mahzam in Qatar; and Uzen and Zhetybay in Kazakhstan.

Some heavy oil fields have been found to be associated with giant gas fields. These include the Bressay, Clair, and Ekofisk gas fields of the North Sea and the Russkoye gas field of Russia.

Table 4 shows world recoverable reserves, undiscovered resources, and total heavy oil endowment by region, as modified from U.S. Geological Survey estimates. Approximately one-half of the world's original heavy oil endowment is thought to have originally existed in the Western Hemisphere. This is due to the enormous heavy oil deposits concentrated in a belt 700 kilometres long by 60 kilometres wide along the Orinoco River in eastern Venezuela.

The Middle East has 36 percent of the world's heavy oil endowment, followed by the United States with 11 percent and Russia with 6 percent. Russia, however, may have large deposits of heavy oil not referenced in available Western technical literature and thus may rank higher in heavy oil resources.

Canadian tar sand deposits

Tar sands. Tar sand deposits occur predominantly in the Western Hemisphere (Table 5, as modified from U.S. Geological Survey estimates). Three-quarters of the total world endowment of bitumen is estimated to occur in the Athabasca region of western Canada. The only two existing commercial bitumen-synthetic oil facilities are located

Table 5: Recoverable Tar Sand (Bitumen) Resources of the World*

region	recoverable resources
United States	4.3
Canada	265.5
South America	0.9
Western Europe	0
Eastern Europe	0.1
Russia	76.2
Transcaucasia and Central Asia	2.0
Africa	4.0
Asia	0
China	1.0
Total world	354.0

*In billion barrels, assuming a 10 percent recovery factor; figures adapted from U.S. Geological Survey, *Oil & Gas Journal*, and American Association of Petroleum Geologists.

in the Athabasca River valley of Alberta. One-fifth of the world's bitumen is thought to be in Russia, mostly in the Volga-Urals and Siberian regions.

Status of the world's supply. The world's total original endowment of recoverable heavy oil has been estimated at about 467,500,000,000 barrels. Of this amount, only 14 percent has been produced (about 67,400,000,000 barrels) and only 20 percent (about 93,000,000,000 barrels) is thought to remain undiscovered. Accordingly, there is a known recoverable reserve of approximately 307,000,000,000 barrels, which can be compared with medium and light oil reserves of 1,119,000,000,000 barrels. As world oil production continues, the future concentration of conventional crude oil will increasingly favour the Middle East. Thus, the heavy oil resources, especially those of the Western Hemisphere and Russia, will continue to be exploited. The Middle East, too, has very significant heavy oil resources and has produced more heavy oil than any other region of the world. Such production also is expected to increase.

The world's total original endowment of recoverable bitumen has been estimated at about 354,000,000,000 barrels. Very little of this has so far been mined for upgrading to syncrude. Hence, the world's total recoverable heavy hydrocarbon component is about 661,000,000,000 barrels, compared with 1,119,000,000,000 barrels of medium and light crude oils. If undiscovered resources are added, the heavy hydrocarbons total 754,000,000,000 barrels and the light and medium oils 1,666,000,000,000 barrels. Some 2,420,000,000,000 barrels of crude oil remain from an original world endowment of 3,211,000,000,000 barrels, and heavy hydrocarbons make up almost one-third of the remaining world crude oil resources.

RECOVERY AND EXPLOITATION

Heavy oils. The recovery of heavy crude oils is impeded by a viscous resistance to flow at reservoir temperatures. The heating of heavy crudes markedly improves their mobility and promotes their recovery. Heat may be introduced into the reservoir by injecting a hot fluid, such as steam or hot water, or by burning some of the heavy oil in the reservoir (a process referred to as in situ combustion or fire flooding).

Heating of heavy oils

Steam soak. A common method involving the use of steam to recover heavy oil is known as steam soak, or steam cycling. It is essentially a well-bore stimulation technique in which steam generated in a boiler at the surface is injected into a production well for a number of weeks, after which the well is closed down for several days before being put back into production. In many cases there is a significant increase in output. It is sometimes economic to steam-soak the same well several times, even though heavy oil recovery usually declines with each succeeding treatment. Steam soaks are economically effective only in thick permeable reservoirs in which vertical (gravity) drainage can occur.

Steam flooding. Continuous steam injection heats a larger portion of the reservoir and achieves the most efficient heavy oil recoveries. Known as steam flooding, this technique is a displacement process similar to waterflooding. Steam is pumped into injection wells, which in some cases are artificially fractured to increase reservoir permeability, and the oil is displaced to production wells. Because of the relatively high cost of steam, water is sometimes injected at an optimum time to push the steam toward the production wells. Because the steam serves two functions, the heating and the transporting of the oil, some steam must always be circulated through the rock formation without condensing. Even in some of the most favourable reservoirs, it is necessary to consume an amount of energy equivalent to burning roughly 25 to 35 percent of the heavy oil produced in order to generate the required amount of steam.

In situ combustion. The mechanics of heavy oil displacement in an in situ combustion operation is similar to that in the steam-flooding process except for one difference. Unlike in the latter, steam is produced by vaporizing water already in the rock formation or water that has been injected therein with heat from the in situ combustion of some of the oil in the reservoir. After the in-place heavy oil has been ignited, the burning front is moved along by continuous air injection. In one variation of the in situ combustion process known as forward combustion, air is injected into a well so as to advance the burning front and heat and displace both the oil and formation water to surrounding production wells. A modified form of forward combustion incorporates the injection of cold water along with air to recover some of the heat that remains behind the combustion front. The air-water combination minimizes the amount of air injected and the amount of in-place oil burned (to between 5 and 10 percent). In another variation of in situ combustion called reverse combustion, a short-term forward burn is initiated by air injection into a well that will eventually produce oil, after which the air injection is switched to adjacent wells. This process is used for recovering extremely viscous oil that will not move through a cold zone ahead of a forward-combustion front.

The costs associated with the generation of heat within a heavy oil reservoir and the success of the recovery process are influenced by the depth of the reservoir. In general, shallower reservoirs are candidates for steam soaks and steam floods, deeper reservoirs for in situ combustion.

Solvent extraction. Solvent extractions also have been used to recover heavy oils. In this process a solvent or emulsifying solution is injected into a heavy oil reservoir. The fluid dissolves or emulsifies the oil as it advances through the permeable reservoir. The oil and fluid are then pumped to the surface through production wells. At the surface the oil is separated from the fluid, and the fluid is recycled.

Tar sands. The bitumen in tar sands can be recovered by surface mining. Open-pit mining methods can be employed where thick deposits occur near the surface. Earth-moving equipment is used to strip and stockpile the topsoil, remove and dispose of the overburden, and excavate the tar sand. The recovery efficiency of surface mining tar sands is estimated at roughly 90 percent. A mill is required to separate the bitumen from the sand in order to upgrade it to commercial quality. This process includes crushing the tar sand and separating the bitumen by mixing the crushed ore with steam and hot water. The bitumen is concentrated by floating and is then treated with a solvent for final separation from the sand and water. The cleaned crude bitumen is upgraded in a delayed coking unit, which produces a blend of lighter hydrocarbon fractions that yield synthetic crude oil, naphtha, kerosene, and gas oil. While there are a large number of heavy oil fields in production throughout the world, only two commercial tar sand surface-mining and synthetic-oil processing operations exist. Both are in western Canada.

Surface mining of tar sands

Economic and technical constraints. Unfortunately, there are problems associated with the exploitation of heavy hydrocarbons. Costs for tar sand mining and upgrading are considerably higher than for producing conventional oil, even in most frontier areas. The tar sand that is mined and milled along with the bitumen is very abrasive and causes rapid wear of equipment. Also, mills and upgrading (coking) facilities are very expensive.

Likewise, the heavy oils are a less desirable energy resource than the lighter crudes, for they are much more costly to extract and to process. An average of about one barrel of oil is combusted (or its energy equivalent expended) to produce the heat necessary to net two barrels of recovered heavy oil. This reduces the recoverable oil in a heavy oil reservoir by one-third.

If the heavy oil is transported by pipeline, direct heating is often required before it will flow at an acceptable rate, necessitating the combustion of additional fuel. The refining of heavy oil results in low yields of distillate products (*e.g.*, naphtha, kerosene, jet fuel, gasoline, oil, and diesel) and correspondingly high yields of sulfur and high-viscosity residues (*e.g.*, asphalt and coke) with metals concentrated in them.

Even under thermal stimulation, heavy oil production ranges only from about 5 to 100 barrels per day per well. This can be compared with recoveries in giant conventional oil fields on the order of 10,000 barrels per day per well. Consequently, even though heavy oil fields are exploited at much slower rates than conventional fields, many more wells are still required. This considerably increases development costs.

Oil shale

Oil shale is a sedimentary rock containing various amounts of solid organic material that yields hydrocarbons, along with a variety of solid products, when subjected to pyrolysis—a treatment that consists of heating the rock to about 500° C. The liquid oil extracted from oil shale as well as that derived from tar sands is referred to as syncrude. Most of the solid by-products of oil shale are unusable wastes, but a few of them have commercial value. These include sulfur, ammonia, alumina, soda ash, and nahcolite (a material that can be used in an industrial air-pollution control procedure known as stack-gas scrubbing).

Syncrude

HISTORY OF USE

Discovery and early application. The first notable reference to oil from shale was in 1596, when the personal physician of Duke Frederick of Württemberg mentioned that mineral oil distilled from oil shale could be used for healing. In 1694, during the reign of William and Mary, British Crown Patent No. 330 was granted to three subjects who had found "a way to extract and make great quantityes of pitch, tarr, and oyle out of a sort of stone." Also about this time, enough oil was actually produced by the distillation of oil shale to light the streets of Modena, Italy.

A commercial shale oil industry was founded in 1838 in Autun, Fr., to produce lamp fuel. By the middle of the 19th century the demand for oil was much greater than could be supplied by the whaling industry. As oil prices rose, numerous oil shale retorts were constructed along the Ohio River in the United States. The first one was built in 1855, but all of them had disappeared by 1860 because E.L. Drake's discovery of crude oil in Pennsylvania in 1859 (see above) opened the market to a cheaper source of oil. Oil shale was retorted in Canada from 1859 to 1861 on the shores of Lake Huron in southwestern Ontario but became economically unattractive with the discovery of crude oil nearby. In Scotland, however, a commercial shale oil industry began in 1862 and operated for about 100 years until the resource was depleted.

A number of other countries also developed oil shale processing facilities: Australia in 1865, Brazil in 1881, New Zealand in 1900, Switzerland in 1915, Sweden in 1921, Spain in 1922, and South Africa in 1935. By 1966, however, all these shale oil plants had closed.

In eastern Europe, oil shale retorting was initiated in Estonia in 1921. The process has been continued to the present, with a daily production of approximately 32,000 barrels of oil. An oil shale processing operation that opened in 1929 in Manchuria in northeastern China also is still producing. It yields an estimated 5,000 barrels of oil per day.

The western oil shales of the United States have been considered economically valuable for more than 70 years. During the mid-1800s, oil was burned and distilled locally from shale in Utah. In Colorado, shale oil was used as smudge in peach orchards about the end of the 19th century. No appreciable output of shale oil, however, was realized until the 1920s, when some 3,600 barrels were produced at a U.S. government plant at Rulison, Colo., and more than 12,000 barrels from a private industrial operation in Nevada. These facilities were closed by 1930 in the wake of the discovery of major conventional oil fields in Texas, Oklahoma, and California.

Modern importance. The shale oil industry reached its highest point of development immediately after World War II. At that time, however, plants were small, with capacities of only about 350 to 1,500 barrels of oil per day. Cost of production was high because of the labour necessary for mining and crushing the rock. During the 1950s the low cost of oil shipped from the Middle East made shale oil uneconomic, and the greatly increased consumption of oil made the amount of shale oil produced insignificant. Therefore, all shale oil plants throughout the world closed down between 1952 and 1966, with the exception of those in Estonia and Manchuria, where mining and economic conditions justified continued exploitation.

Competition with conventional oil

The cycles of competition between synthetic and conventional oil are related to discoveries of major conventional oil fields. The giant oil fields found in the Middle East after World War II virtually eliminated any remaining commercial interest in the production of shale oil until the energy shortages of the late 1960s and '70s. These shortages prompted several countries to survey their oil shale deposits in order to determine whether sufficient reserves were present to justify the large investments that would be needed to turn shale oil into a practical energy source.

Brazil and the United States developed pilot plants with which to exploit their deposits of oil shale: the Irati shales of Permian age in Brazil and the Green River shales of Eocene age (about 36,600,000 to 57,800,000 years old) in Colorado, Utah, and Wyoming. These two oil shale deposits are the largest in the world. The Brazilian national oil company, Petróleo Brasileiro (commonly called Petrobrás), developed a commercial shale oil extraction technology on the basis of work at a pilot plant that began operation since 1972. The operation was located at São Mateus do Sul in Paraná in southern Brazil. More than 1,500,000 barrels of shale oil and 20,000 tons of sulfur were extracted from over 3,500,000 tons of Irati oil shale before the project was canceled owing to the relatively low costs of conventional oil after the mid-1980s.

Since the early 1980s, shale oil development in the United States has also been limited in large part because of falling oil prices resulting from increased world crude oil production. The only completed plant capable of producing shale oil is the Unishale B retort in Parachute Creek, Colo. The plant is still in the testing phase, and only experimental amounts of shale oil have been produced. It has a projected capacity of about 10,400 barrels per day, however. In the United States, as in most other countries, shale oil development faces an uncertain future, at least until conventional oil resources are more nearly depleted.

(J.P.Ri./G.I.A.)

COMPOSITION, FORMATION, AND OCCURRENCE

Mineral and organic constituents. The mineral constituents of oil shales vary according to sediment type. Some are true shales in which clay minerals are predominant. Others, such as the Green River shales of the western United States, are carbonates (*e.g.*, dolomites and limestones) with subordinate amounts of other minerals. The various oil shales that have been mined during the past century have ranged from shale to marls and carbonates. The excluded sedimentary variety is sandstone, because the environment in which sandstones are deposited is not compatible with the accumulation and preservation of organic matter.

Kerogen

The organic matter contained in oil shales is principally kerogen, an insoluble solid material. The shales range from brown to black in colour. They have low specific weight and are flammable, burning with a sooty flame. Their external structure is laminar, and, in a stratigraphic section, alternating darker and lighter strata correspond to the periodic changes of organic content. Oil shales are quite resistant to the oxidizing effects of air. In terms of chemical composition, they consist primarily of silica, iron, aluminum, calcium, magnesium, sodium carbonates, silicates, oxides, and sulfides. The chemical composition of the organic matter in oil shales is variable, but the hydrogen content is always high. The oxygen content varies, as does the amount of nitrogen, which is much less abundant.

Formation. Some oil shale kerogens are composed almost entirely of algal remains, whereas others are a mixture of amorphous organic matter with a variable content of identifiable organic remnants. The main algal types are *Botryococcus* and *Tasmanites.*

Botryococcus is a fresh- or brackish-water alga that forms colonies. Permian kerogens from Autun, Fr., and Carboniferous and Permian torbanite from Scotland, Australia, and South Africa appear to consist almost exclusively of

Botryococcus colonies, as does Recent (post-Pleistocene) coorongite from Australia.

Tasmanites is a marine alga the remains of which make up nearly all the kerogen of such oil shales as the Permian tasmanite of Australia and the Jurassic-Cretaceous tasmanite of Alaska. The remains of *Tasmanites* also are present in many other shales, such as the Lower Toarcian shales (those about 190,000,000 years in age) of the Paris Basin in France and the Lower Silurian shales (those about 423,000,000 years in age) of Algeria.

Often only a minor part of the kerogen in oil shales is made of recognizable organic remnants. The rest is amorphous, probably because of microbial alteration during sedimentation. Amorphous organic material (sapropelic matter) associated with minerals constitutes thick accumulations of oil shale, such as the Permian Irati shales of Brazil and the Eocene Green River shales of the western United States. The organic material may have been derived from planktonic organisms (*e.g.*, algae, copepods, and ostracods) and from microorganisms (*e.g.*, bacteria and algae) that normally live in fresh sediment.

A characteristic typical of the various types of oil shale is a very fine lamination of thin alternating layers of minerals and organic matter. This lamination results from sedimentation in quiet waters in which either carbonates are precipitated from solution or clay minerals are transported as extremely fine detritus. Also, a succession of seasonal or other periodic events is suggested by the layering.

Green River shales

The geologic environment. A common geologic environment in which oil shales, often of considerable thickness, are deposited is large lake basins, particularly those of tectonic origin. Mineralogically, these oil shales are marls or argillaceous limestones, which may be associated with volcanic tuffs and evaporites. The major oil shale deposits of this type are the Green River shales of Eocene age in the western United States, along with the oil shales of Triassic age (about 208,000,000 to 245,000,000 years old) in Zaire and the Albert shales of Mississippian origin (roughly 320,000,000 to 360,000,000 years old) in New Brunswick, Can.

Oil shales deposited in shallow marine environments are thinner but of greater areal extent. The mineral phase is mostly clay and silica minerals, though carbonates also may occur. Extensive deposits of black shales of this variety were formed during the Cambrian Period (from about 505,000,000 to 540,000,000 years ago) in northern Europe and Siberia; the Silurian (about 408,000,000 to 438,000,000 years ago) in North America; the Permian (about 245,000,000 to 286,000,000 years ago) in southern Brazil, Uruguay, and Argentina; the Jurassic (about 144,000,000 to 208,000,000 years ago) in western Europe; and the Miocene Epoch of the Tertiary (about 5,300,000 to 23,700,000 years ago) in Italy, Sicily, and California.

Oil shales also have been deposited in small lakes, bogs, and lagoons where they are associated with coal seams. Deposits of this type occur in the Permian sequence of western Europe and in the Tertiary beds of Manchuria (Northeast), China.

WORLD OIL SHALE RESOURCES

Oil shales are found in many places throughout the world, yet worldwide shale oil development has been economically more attractive than conventional oil development for only a few brief periods in the 20th century. In earlier centuries oil shales were successfully exploited in a number of locations (see above).

The United States. Oil shale deposits that are commercially viable at present are the Estonian shales and the Manchurian shales in China. Nearly 60 percent of the world's potentially recoverable shale oil resource is concentrated, however, in the United States (Table 6). The aforementioned western and minable eastern oil shales of the United States have been estimated to contain an in-place oil resource of some 1,670,000,000,000 barrels. Using a 50 percent allowance for unrecoverable shale and a 25 percent allowance for conversion to synthetic fuel, the production potential for shale oil in the United States is estimated to be 626,000,000,000 barrels. The Mahogany Zone of the Parachute Creek Member of the Eocene Green River formation in the Piceance Basin of northwestern Colorado is a major target for future shale oil production. Estimated to contain some 59,000,000,000 barrels of recoverable shale oil, it is a thick, rich, consistent, saucer-shaped bed that outcrops around the edges of the basin, offering opportunities for mining by adits (nearly horizontal passages from the surface). At the centre of the basin the zone is more than 150 metres deep and accessible only by vertical or inclined shafts.

Piceance Basin, Colorado

Table 6: The Recoverable Shale Oil Resources of the United States*

deposits	recoverable resources†
Piceance Basin (Colorado)	
Mahogany Zone	59
Shales above Mahogany Zone	90
Shales below Mahogany Zone	231
Uinta Basin (Utah)	51
Other western basins	131
Eastern oil shales (Kentucky, Indiana, Ohio)	64
Total	626

*Recovery factor = 37.5 percent of estimated in-place resource. †In billion barrels; figures adapted from *Oil & Gas Journal*, U.S. Geological Survey, and American Association of Petroleum Geologists.

The western oil shales of the United States are richer than its eastern shales, yielding from 84 to 168 litres of raw shale oil for each metric ton of oil shale processed (from 20 to 40 U.S. gallons per short ton). The oil is relatively high in paraffins. Thus, with upgrading, it becomes an excellent refinery feedstock that is well suited to large yields of diesel and jet fuel.

Although the organic content of western and eastern oil shales is the same, the eastern shales yield only 34 to 63 litres of raw oil per metric ton of oil shale. Eastern shale oils are more aromatic and, when upgraded, are better suited as a feed for catalytic crackers in the production of gasoline.

Worldwide. Table 7 provides estimates of the world's recoverable shale oil as reported in the technical literature. Allowance is made for unrecoverable shale and for conversion to synthetic fuel. Brazil's oil shale resources are the world's second largest (28 percent of the total). Estonia, Russia, Zaire, Australia, Canada, Italy, and China also have significant oil shale resources. The world's total recoverable shale oil resource is estimated at some 1,067,000,000,000 barrels.

RECOVERY AND EXPLOITATION

Minimum organic requirement. As noted above, the organic matter in oil shales is kerogen, with no oil and little extractable bitumen naturally present. The kerogen of oil shale is not distinct from the kerogen of petroleum source rocks, and to some extent the pyrolysis process for extracting oil from oil shales is comparable to the burial of source rocks at depth and the subsequent formation of oil by the resulting elevation of temperature.

Nonetheless, oil shale must have a large amount of organic matter to be of commercial interest, larger than the

Table 7: The Recoverable Shale Oil Resources of the World*

region	recoverable resources†
United States	626
Canada	16
South America (Brazil)	300
Northern and western Europe	2
Italy	13
Estonia	1
Russia	41
Zaire	38
China	10
Australia	17
Other countries	3
Total world	1,067

*Recovery factor = 37.5 percent of estimated in-place resource. †In billion barrels; figures adapted from *Oil & Gas Journal*, U.S. Geological Survey, and American Association of Petroleum Geologists.

0.5 percent of organic carbon in a source rock from which commercial accumulations of oil or gas may be generated, provided that depth of burial, migration paths, and trapping mechanisms are favourable. The organic matter in a commercial oil shale must provide more energy than is required to process the shale. If the kerogen content of the shale is 2.5 percent by weight, its total calorific value is needed for processing. This is because at an average pyrolysis temperature of 500° C the energy required for heating is about 250 calories per gram of rock and the calorific value of kerogen is 10,000 calories per gram. Oil shale with a kerogen content below the threshold of 2.5 percent therefore cannot be employed as a source of energy. Frequently used is a lower limit of 5 percent organic content, which corresponds to an oil yield of approximately 25 litres per metric ton (6 U.S. gallons per short ton) of rock. Even this amount is not considered of potentially commercial grade in the United States, where 10 U.S. gallons per short ton (42 litres per metric ton) is often cited as a lower limit for oil shales.

Oil yield per ton

Recovery processes. *Pyrolysis.* The technology for producing oil from oil shale is based on pyrolysis of the rock. The heat breaks the various chemical bonds of the kerogen macromolecule, liberating small molecules of liquid and gaseous hydrocarbons, as well as nitrogen, sulfur, and oxygen compounds. However, since the shale quickly reaches a high temperature, industrial reactions are somewhat different from those that occur under natural subsurface conditions. Thus, liquid industrial products usually include a large proportion of olefins and sulfur and nitrogen compounds. Also, industrial gases may contain large amounts of hydrogen sulfide and ammonia.

Aboveground processing. Three basic steps are involved in the aboveground processing of oil shales—mining, crushing, and retorting (heating). Various retorting processes have been used over the years. The Pumpherston process, which involves external heating through the wall of the retort, was used in Scotland beginning in 1862. This process was widely employed with various refinements introduced later in continental Europe. The capacities of the retorting units, however, were low and energetic balances poor.

Combustion inside the retorting unit results in better energetic balances, but low-calorific gas diluted by nitrogen and combustion products results. This technique is used in Russia and China and is being tested in the United States. Still another method involves the circulation of externally heated gas through the shale. The resulting energetic balance is satisfactory, and the gas produced is of high calorific value. Used in France, this approach was adopted by Brazil and is the subject of experimentation in the United States.

An entirely new experimental process involves heating the shale with hot solids, which ensures a good energetic balance and a high calorific value of the gas produced. In certain parts of the world hot shale ashes are used as a calorific vehicle, while in the United States externally heated ceramic balls are employed. This technology is more complex than any of the others.

Subsurface processing. Subsurface processing differs from aboveground processing in that retorting to produce oil and gas takes place underground, or in situ. The oil shale is fractured underground by explosives and then heated by a controlled underground fire. Fuels produced from the heated oil shale are pumped to the surface and collected. Several in situ processes have been tested in the United States; they have resulted in both high and low rates of recovery efficiency.

Economic and technical constraints. An emerging shale oil industry faces many economic uncertainties, the most significant of which is the comparatively lower price of conventional oil development. The capital requirements for a commercial shale oil project range up to several billion U.S. dollars, making almost any conventional oil development, with comparable production, less expensive. Thus, it is unlikely that wide-scale exploitation of shale oil will occur until conventional and heavy oils are more nearly depleted.

High capital costs

Furthermore, technical difficulties remain in retorting. In the United States, for example, component design problems in the retorting process have hampered the operation of a commercial-scale, aboveground facility. Also, experiments with in situ retorting have resulted in quite varied recovery efficiencies.

In addition, the shale oil industry may be constrained by environmental factors. In semiarid regions, such as the western United States, limited water resources pose a problem because large amounts of water are required for the extraction process. Moreover, mining and processing may have adverse effects on groundwater and air quality. The vast amounts of rock material that have to be moved in a shale oil recovery operation also may adversely affect the integrity of the land, grazing and agricultural activities, and local fauna and flora.

There is little prospect of widespread oil shale exploitation in the near term. Such broad development must await more favourable economic conditions that would most likely be brought about by the depletion of conventional and heavy oil resources. Even then technological and environmental constraints would have to be dealt with. Oil shale could possibly become a major supplier of the world's energy but probably not until well into the 21st century. (J.P.Ri.)

Coal

Coal is a combustible rock that contains more than 50 percent by weight (or 70 percent by volume) carbonaceous matter produced by the compaction and induration of altered plant remains—namely, peat deposits. Different varieties of coal arise because of differences in the kinds of plant material (coal type), degree of coalification (coal rank), and range of impurities (coal grade). Coals are rich in carbon and are usually brown or black in colour. Most occur in stratified, sedimentary deposits; however, these deposits may later be subjected to elevated temperatures and pressures due to igneous intrusions or deformation during orogenesis (*i.e.*, processes of mountain building), resulting in the development of anthracite and even graphite. Although the concentration of carbon in the Earth's crust does not exceed 0.1 percent by weight, it is indispensable to life and constitutes humankind's main source of energy.

Variations among coals

HISTORY OF USE

Discovery and early application. *In ancient times.* The discovery of the use of fire helped to distinguish humans from other animals. Early fuels were primarily wood (and charcoal derived from it), straw, and dried dung. References to the early uses of coal are meagre. Aristotle referred to "bodies which have more of earth than of smoke" and called them "coal-like substances." (It should be noted that biblical references to coal are to charcoal rather than to the rock, coal.) Coal was used commercially by the Chinese long before it was utilized in Europe. Although no authentic record is available, coal from the Fu-shun mine in northeastern China may have been employed to smelt copper as early as 1000 BC. Cast-iron Chinese coins dating to about the 1st century BC are thought to have been made using coal. Stones used as fuel were said to have been produced in China during the Han dynasty (206 BC–AD 220).

In Europe. Coal cinders found among Roman ruins in England suggest that the Romans were familiar with its use before AD 400. The first documented proof that coal was mined in Europe was provided by the monk Reinier of Liège, who wrote (about 1200) of black earth very similar to charcoal used by metalworkers. Many references to coal mining in England, Scotland, and the European continent began to appear in the writings of the 13th century. Coal was, however, used only on a limited scale until the early 18th century when Abraham Darby of England and others developed methods of using coke made from coal in blast furnaces and forges. Successive metallurgical and engineering developments—most notably the invention of the coal-burning steam engine by James Watt—engendered an almost insatiable demand for coal.

Coke

In the New World. Up to the time of the American

Revolution, most coal used in the American colonies came from England or Nova Scotia. Wartime shortages and the needs of the munitions manufacturers, however, spurred on small American coal-mining operations such as those in Virginia on the James River near Richmond. By the early 1830s mining companies had emerged along the Ohio, Illinois, and Mississippi rivers and in the Appalachian regions. As in European countries, the introduction of the steam locomotive gave the American coal industry a tremendous impetus. Continued expansion of industrial activity in the United States and in Europe further promoted the use of coal.

Modern utilization. *As an energy source.* Coal is an abundant natural resource that can be used as a source of energy, as a chemical feedstock from which numerous synthetic compounds (*e.g.*, dyes, oils, waxes, pharmaceuticals, and pesticides) can be derived, and in the production of coke for metallurgical processes. Coal is presently a major source of energy in the production of electrical power using steam generation. In addition, gasification and liquefaction produce gaseous and liquid fuels that can be easily transported (*e.g.*, by pipeline) and conveniently stored in tanks.

Conversion. In general, coal can be considered a hydrogen-deficient hydrocarbon with a hydrogen-to-carbon ratio near 0.8, as compared with a liquid hydrocarbons ratio near 2 and a gaseous hydrocarbons ratio near 4. For this reason, any process used to convert coal to alternative fuels must add hydrogen (either directly or in the form of water). Gasification refers to the conversion of coal to a mixture of gases, including carbon monoxide, hydrogen, methane, and other hydrocarbons, depending on the conditions involved. Gasification may be accomplished either in situ or in processing plants. In situ gasification is accomplished by controlled, incomplete burning of a coal bed underground while adding air and steam. The gases are withdrawn and may be burned to produce heat, generate electricity, or be used as synthesis gas in indirect liquefaction or the production of chemicals.

Liquefaction may be either direct or indirect (*i.e.*, by using the gaseous products obtained by breaking down the chemical structure of coal). Four general methods are used for liquefaction: (1) pyrolysis and hydrocarbonization (coal is heated in the absence of air or in a stream of hydrogen), (2) solvent extraction (coal hydrocarbons are selectively dissolved and hydrogen is added to produce the desired liquids), (3) catalytic liquefaction (hydrogenation takes place in the presence of a catalyst—for example, zinc chloride), and (4) indirect liquefaction (carbon monoxide and hydrogen are combined in the presence of a catalyst).

Problems associated with the use of coal. *Hazards of mining and preparation.* Coal is very abundant; estimates of reserves indicate that enough coal remains to last for many hundreds of years. There is, however, a variety of problems associated with the use of coal. Mining operations are hazardous. Each year hundreds of coal miners lose their lives or are seriously injured. Major mine hazards include roof falls, rock bursts, and fires and explosions. The latter result when flammable gases (such as methane) trapped in the coal are released during mining operations and accidentally are ignited. Promising research in the extraction of methane from coal beds prior to mining is expected to lead to safer mines and provide a source of natural gas that has been wasted for so long. Also, the repeated inhalation of coal dust over extended periods of time can result in serious health problems, as, for example, anthracosis (commonly called black lung disease).

Environmental problems

Coal mines and coal-preparation plants caused much environmental damage in the past (Figure 8). Surface areas exposed during mining, as well as coal and rock waste (which were often dumped indiscriminately), weathered rapidly, producing abundant sediment and soluble

From *The Direct Use of Coal*, Prospects and Problems of Production and Combustion, 1979, Office of Technology Assessment, Washington, D.C.

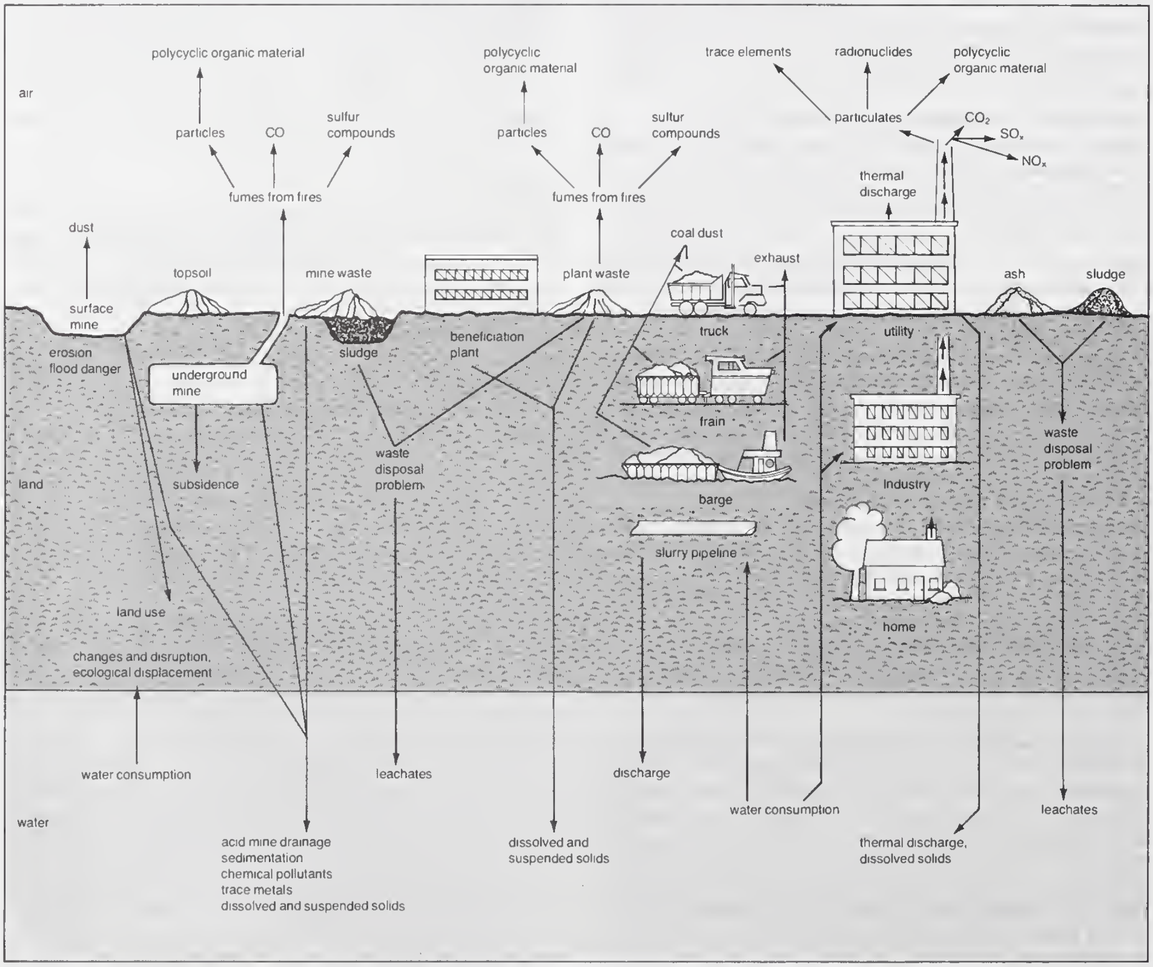


Figure 8: Environmental disturbances from coal-related activities.

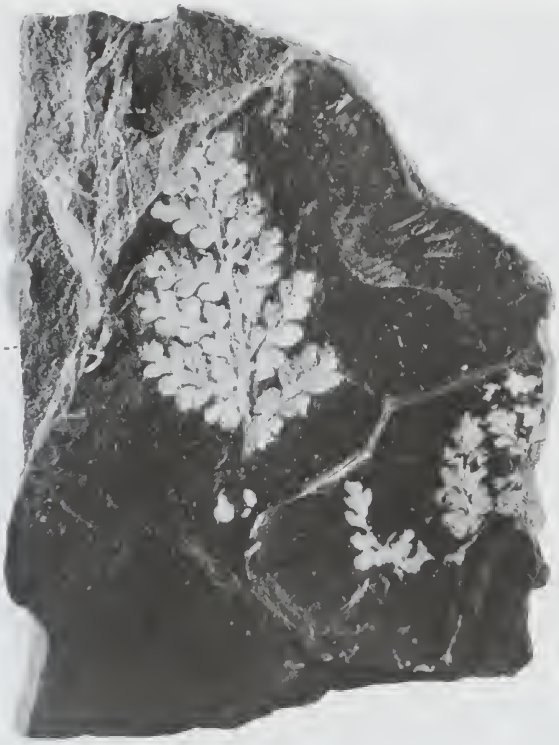


Figure 9: Lignite coal with fern fossilization.
Runk/Schoenberger—Grant Heilman

chemical products such as sulfuric acid and iron sulfates. Nearby streams became clogged with sediment, iron oxides stained rocks, and "acid mine drainage" caused marked reductions in the numbers of plants and animals living in the vicinity. Potentially toxic elements, leached from the exposed coal and adjacent rocks, were released into the environment. Since the 1970s, however, stricter laws have significantly reduced the environmental damage caused by coal mining.

Hazards of utilization. Coal utilization also can cause problems. During the incomplete burning or conversion of coal, many compounds are produced, some of which are carcinogenic. The burning of coal also produces sulfur and nitrogen oxides that react with atmospheric moisture to produce sulfuric and nitric acids—so-called acid rain. In addition, it produces particulate matter (fly ash) that can be transported by winds for many hundreds of kilometres and solids (bottom ash and slag) that must be disposed of. Trace elements originally present in the coal may escape as volatiles (*e.g.*, chlorine and mercury) or be concentrated in the ash (*e.g.*, arsenic and barium). Some of these pollutants can be trapped by using such devices as electrostatic precipitators, baghouses, and scrubbers. Current research on alternative means for combustion (*e.g.*, fluidized bed combustion, magnetohydrodynamics, and low nitrogen dioxide burners) is expected to provide efficient and environmentally attractive methods for extracting energy from coal. Regardless of the means used for combustion, acceptable ways of disposing of the waste products have to be found. Finally, the burning of all fossil fuels (oil and natural gas included) releases large quantities of carbon dioxide (CO_2) into the atmosphere. The CO_2 molecules allow the shorter-wavelength rays from the Sun to enter the atmosphere and strike the Earth's surface, but they do not allow much of the long-wave radiation reradiated from the surface to escape into space. The CO_2 absorbs this upward-propagating infrared radiation and reemits a portion of it downward, causing the lower atmosphere to remain warmer than it would otherwise be. This is the so-called greenhouse effect. (Other gases, such as methane and ozone, are also greenhouse gases.) Whether higher concentrations of CO_2 in the atmosphere will in fact lead to significant climatic changes has not yet been clearly established, but it is a matter of considerable concern among scientists. A rise in the Earth's mean annual temperature of only a few degrees Celsius could cause melting of enough glacial ice to inundate coastal areas. Technologies being considered to reduce carbon dioxide levels include biological fixation, cryogenic recovery, disposal in the oceans and aquifers, and conversion to methanol.

ORIGIN

Coal-forming materials. *Plant matter.* It is generally accepted that most coals formed from plants that grew in and adjacent to swamps in warm, humid regions. Material derived from these plants accumulated in low-lying areas that remained wet most of the time and was converted to peat through the activity of microorganisms. (It should be noted that peat can occur in temperate regions [*e.g.*, Ireland and the state of Michigan in the United States] and even in subarctic regions [*e.g.*, the Scandinavian countries].) Under certain conditions this organic material continued to accumulate and was later converted into coal. Much of the plant matter that accumulates on the surface of the Earth is never converted to peat or to coal, because it is removed by fire or organic decomposition. Hence, the vast coal deposits found in ancient rocks must represent periods during which several favourable biological and physical processes occurred at the same time.

Evidence of plant origin

Evidence that coal was derived from plants comes from three principal sources. First, lignites, the lowest coal rank, often contain recognizable plant remains (Figure 9). Second, sedimentary rock layers above, below, and adjacent to coal seams contain plant fossils in the form of impressions and carbonized films (*e.g.*, leaves and stems) and casts of larger parts such as roots, branches, and trunks. Third, even coals of advanced rank may reveal the presence of precursor plant material. When examined microscopically in thin sections or polished blocks, cell walls, cuticles (the outer wall of leaves), spores, and other structures can still be recognized (see below *Genesis of macerals and rock types*). Algal and fungal remains also may be present. (Algae are major components in boghead coal, which appears to be transitional between coal and petroleum.)

The fossil record. Anthracite (the highest coal rank) material, which appears to have been derived from plants, is known from the Proterozoic of Precambrian times (beginning approximately 2,500,000,000 years ago). Siliceous rocks of the same age contain fossil algae and fungi. These early plants were primarily Protista (solitary or aggregate unicellular organisms that include yellow-green algae, golden-brown algae, and diatoms), which lived in aqueous environments. It was not until the Late Silurian Period (from 408,000,000 to 421,000,000 years ago) that plants are known to have developed the ability to survive on land. Fossil organisms that are reflective of this dramatic evolutionary event have been discovered in Wales and Australia.

Evidence for early coastal forests is preserved in strata of Late Devonian age (about 360,000,000 to 374,000,000 years old). By the latter half of the Paleozoic Era, plants had undergone extensive evolution and occupied many previously vacant environments (this phenomenon is sometimes called adaptive radiation).

Major eras of coal formation

There were two major eras of coal formation in geologic history. The older includes the Carboniferous Period (sometimes divided into the Mississippian and Pennsylvanian periods, from approximately 286,000,000 to 360,000,000 years ago) and the Permian Period (from about 245,000,000 to 286,000,000 years ago). Much of the bituminous coal of eastern North America and Europe is Carboniferous in age. Most coals in Siberia, eastern Asia, and Australia are of Permian origin. The younger era began about 135,000,000 years ago during the Cretaceous Period and reached its peak during the Tertiary Period (about 1,600,000 to 66,400,000 years ago) of the Cenozoic. Most of the coals that formed during this second era are lignites and subbituminous (or brown) coals. These are widespread in such areas as western North America (including Alaska), southern France and central Europe, Japan, and Indonesia.

Late Paleozoic flora included sphenopsids, lycopsids, pteropsids, and Cordaitales. The sphenopsid, *Calamites,* grew as trees in swamps (see Figure 10). *Calamites* had

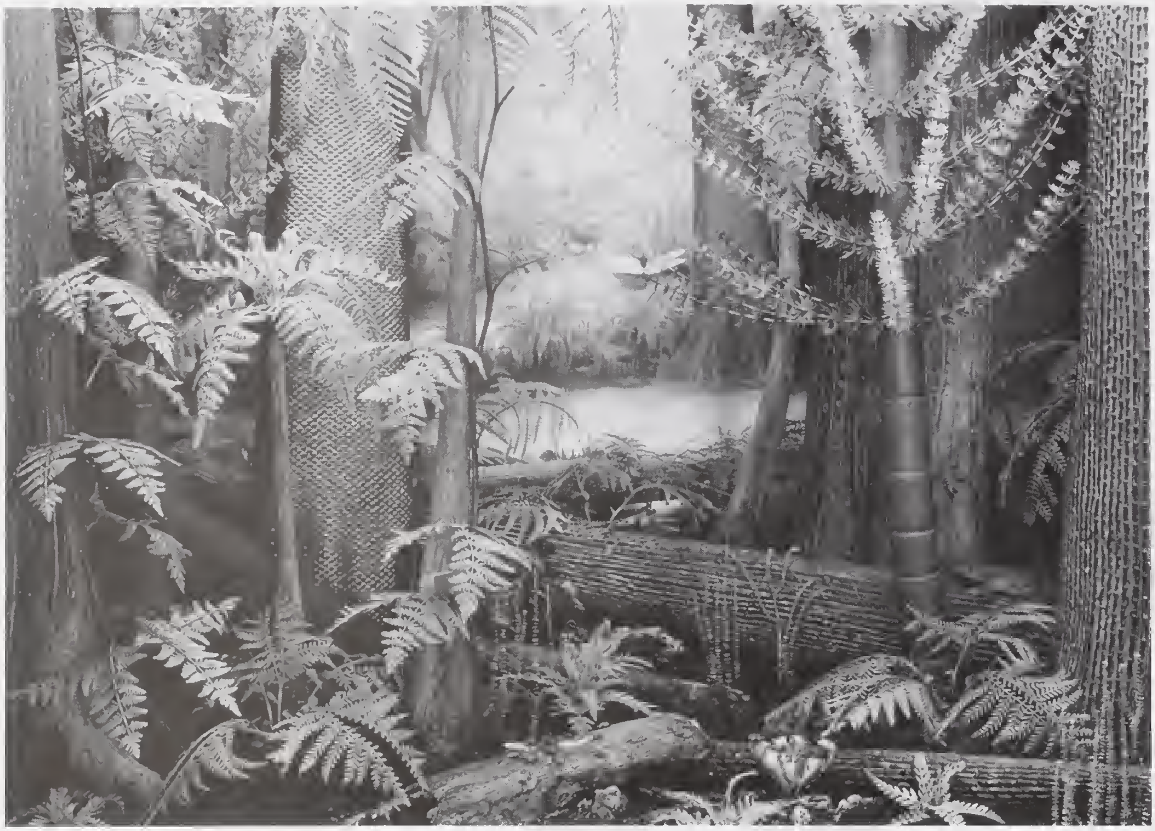


Figure 10: *Pennsylvanian coal forest diorama.*
The lone tree (*Calamites*) with horizontal grooves in the right foreground is a jointed sphenopsid; the large trees with scar patterns are lycopsids.
By courtesy of the Department Library Services, American Museum of Natural History, neg #333983

long, jointed stems with sparse foliage. The lycopsids included species of *Lepidodendron* and *Sigillaria* (up to 30 metres tall) that grew in somewhat drier areas. Pteropsids included both true ferns (Filicineae) and extinct seed ferns (Pteridospermaphyta), which grew in relatively dry environments. The Cordaitales, which had tall stems and long, narrow, palmlike leaves, also favoured drier areas. During the Cretaceous and Cenozoic the angiosperms (flowering plants) evolved, producing a diversified flora from which the younger coals developed.

Formation processes. *Peat.* Although peat is used as a source of energy, it is not usually considered a coal. It is the precursor material from which coals are derived, and the process by which peat is formed is studied in existing swamps in many parts of the world (*e.g.,* the Okefenokee Swamp of Georgia). The formation of peat is controlled by several factors including (1) the evolutionary development of plant life, (2) the climatic conditions (warm enough to sustain plant growth and wet enough to permit the partial decomposition of the plant material and preserve the peat), and (3) the physical conditions of the area (its geographic position relative to the sea or other bodies of water, rates of subsidence or uplift, and so forth). Warm, moist climates are thought to produce broad bands of bright coal. Cooler, temperate climates, on the other hand, are thought to produce detrital coal with relatively little bright coal.

Conditions for peat formation

Initially, the area on which a future coal seam may be developed must be uplifted so that plant growth can be established. Areas near seacoasts or low-lying areas near streams stay moist enough for peat to form, but elevated swamps (some bogs and moors) can produce peat only if the annual precipitation exceeds annual evaporation and little percolation or drainage occurs. Thick peat deposits necessary for coal formation develop at sites where the following conditions exist: slow, continuous subsidence; the presence of such natural structures as levees, beaches, and bars that give protection from frequent inundation; and a restricted supply of incoming sediments that would interrupt peat formation. In such areas the water may become quite stagnant (except for a few rivers traversing the swamp), and plant material can continue to accumulate. Microorganisms attack the plant material and convert it to peat. Very close to the surface where oxygen is still readily available (aerobic, or oxidizing, conditions), the decomposition of the plant material produces mostly gaseous and liquid products. With increasing depth, however, the conditions become increasingly anaerobic (reducing), and molds and peats develop. The process of peat formation—biochemical coalification—is most active in the upper few metres of a peat deposit. Fungi are not found below about 0.5 metre, and most forms of microbial life are eliminated at depths below about 10 metres. If the rate of subsidence and/or the rate of influx of new sediment increases,

Table 8: Petrologic Components (Macerals) in Coal and Their Groupings

maceral grouping in Europe	macerals or components		maceral grouping (constituents) in the United States
	name in Europe*	name in the United States†	
Vitrinite	telinite	megascopic anthraxylon attrital anthraxylon	anthraxylon
	collinite	subanthraxylon humic matter light-brown matter	translucent attritus
Exinite	resinite	red resins yellow resins	
	cerinite	amorphous wax	
	sporinite (exinite)	spore coats	
	cutinite	cuticles	
	suberinite	suberin	
	alginite	algal bodies	
Inertinite	massive micrinite	dark-brown matter amorphous opaque matter	opaque attritus
	granular micrinite	granular opaque matter	
	sclerotinite	fusinized fungal matter	petrologic fusain
	semifusinite	dark semifusain	
	fusinite	attrital fusain megascopic fusain	

*The majority of these names originated with M.C. Stopes (1935) and were adopted by the International Stratigraphical Congresses (1935 and 1951) at Heerlen.
†These names are mainly from R. Thiessen.

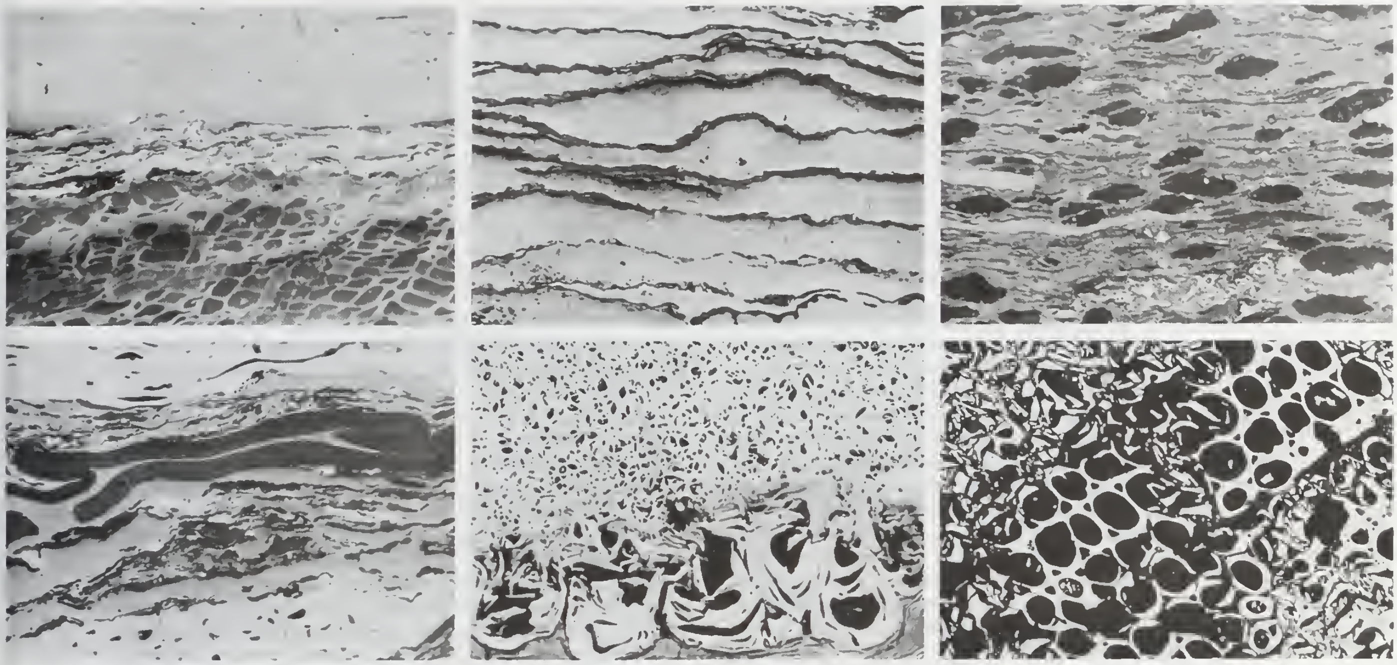


Figure 11: *Typical macerals.*
(Top left) Collinite (upper part) and telinite (lower) with resinite inclusions. (Top centre) Cutinite embedded in collinite. (Top right) Boghead-cannel coal with alginite. (Bottom left) Sporinite (macrospores and microspores) embedded in collinite (gray) and surrounded by micrinite (white). (Bottom centre) Semifusinite (upper part) and sclerotinite (lower). (Bottom right) Fusinite (bogen texture manifested at the upper left and lower right corners). Photomicrographs in reflected light (oil-immersion); magnified about 144×.

By courtesy of M Th Mackowsky, Bergbauforschung Essen Germany

the peat will be buried and soon thereafter the coalification process—geochemical coalification—begins. The cycle may be repeated many times, which accounts for the numerous coal seams found in some sedimentary basins.

Coalification. The general sequence of coalification is from lignite to subbituminous to bituminous to anthracite (see below *Ranking by coalification*). Since microbial activity ceases within a few metres of the Earth's surface, the coalification process must be controlled primarily by changes in physical conditions that take place with depth. Some coal characteristics were determined by events that occurred during peat formation—*e.g.*, charcoallike material in coal is attributed to fires that occurred during dry periods while peat was still forming.

Effects of heat and pressure over time

Three major physical factors—duration, increasing temperature, and increasing pressure—may influence the coalification process. In laboratory experiments artificially prepared coals are influenced by the duration of the experiment, but in nature the length of time is substantially longer and the overall effect of time remains undetermined. Low-rank (*i.e.*, brown) coal in the Moscow Basin was deposited during Carboniferous time but was not buried deeply and never reached a higher rank. The most widely accepted explanation is that coalification takes place in response to increasing temperature. In general, temperature increases with depth. This geothermal gradient averages about 30° C per kilometre, but the gradient ranges from less than 10° C per kilometre in regions undergoing very rapid subsidence to more than 100° C per kilometre in areas of igneous activity. Measurements of thicknesses of sedimentary cover and corresponding coal ranks suggest that temperatures lower than 200° C were sufficient to produce coal of anthracite rank. The effect of increasing pressure due to depth of burial is not considered to cause coalification. In fact, increasing overburden pressure might have the opposite effect if such volatile compounds as methane that must escape during coalification are retained. Pressure may influence the porosity and moisture content of coal.

COAL TYPES AND RANKS

Coals may be classified in several ways. One mode of classification is by coal type; such types have some genetic implications because they are based on the organic materials present and the coalification processes that produced the coal. The most useful and widely applied coal classification schemes are those based on the degree to which coals have undergone coalification. Such varying degrees of coalification are generally called coal ranks (or classes). In addition to the scientific value of classification schemes of this kind, the determination of rank has a number of practical applications. Many coal properties are in part determined by rank, including the amount of heat produced during combustion, the amount of gaseous products released upon heating, and the suitability of the coals for liquefaction or for producing coke.

Coal types. *Macerals.* Coals contain both organic and inorganic phases. The latter consist either of minerals such as quartz and clays that may have been brought in by flowing water (or wind activity) or of minerals such as pyrite and marcasite that formed in place (authigenic). Some formed in living plant tissues, while others formed later during peat formation or coalification. Some pyrite (and marcasite) is present in micrometre-sized spheroids called framboids (from their raspberry-like shape) that formed quite early. Framboids are very difficult to remove by conventional coal-cleaning processes.

By analogy to the term mineral, Marie C. Stopes of Great Britain (1935) proposed the term maceral to describe organic constituents present in coals. The word is derived from the Latin *macerare,* meaning "to macerate." (Mineral names often end in "-ite." The corresponding ending for macerals is "-inite.") Maceral nomenclature has been applied differently by some European coal petrologists who studied polished blocks of coal using reflected-light microscopy (their terminology is based on morphology, botanical affinity, and mode of occurrence) and by some North American petrologists who studied very thin slices (thin sections) of coal using transmitted-light microscopy. Various nomenclature systems have been used. A comparison of common coal maceral terms is given in Table 8, and typical macerals are shown in Figure 11.

Vitrinites, exinites, and inertinites

Three major maceral groups are generally recognized: vitrinite, exinite, and inertinite. The vitrinite group is the most abundant, constituting as much as 50 to 90 percent of many North American coals. Vitrinites are derived

Table 9: Coal Type According to Appearance and Composition

type	main components	opaque attritus percent
Banded coal (>5 percent anthraxylon)		
Bright-banded	anthraxylon and translucent attritus	< 20
Semisplint	translucent and opaque attritus	20–30
Splint	opaque attritus	> 30
Nonbanded coal (<5 percent anthraxylon)		
Cannel	attritus with spores	
Boghead	attritus with algae	

primarily from cell walls and woody tissues. They show a wide range of reflectance values (discussed below), but in individual samples these values tend to be intermediate compared with those of the other maceral groups. Several varieties are recognized—*e.g.*, telinite (the brighter parts of vitrinite that make up cell walls) and collinite (clear vitrinite that occupies the spaces between cell walls).

The exinite (or liptinite) group makes up 5 to 15 percent of many coals. Exinites are derived from waxy or resinous plant parts, such as cuticles, spores, and wound resins. Their reflectance values are usually the lowest in an individual sample. Several varieties are recognized, including sporinite (spores typically preserved as flattened spheroids), cutinite (part of cross sections of leaves, often with crenulated surfaces), and resinite (ovoid and sometimes translucent masses of resin). The exinites may fluoresce under ultraviolet light, but with increasing rank their optical properties approach those of the vitrinites, and they become indistinguishable.

The inertinite group makes up from 5 to 40 percent of most coals. Their reflectance values are usually the highest in a given sample. The most common inertinite maceral is fusinite, which has a charcoallike appearance with obvious cell texture. The cells may be either empty or filled with mineral matter, and the cell walls may have been crushed during compaction (bogen texture; Figure 11). Inertinites are strongly altered or degraded plant material that is thought to have been produced during the formation of peat; for example, charcoal produced by a fire in a peat swamp is preserved as fusinite.

Coal rock types. Coals may be classified on the basis of their macroscopic appearance (generally referred to as coal rock type, lithotype, or kohlentype). Four main types are recognized: (1) vitrain (*Glanzkohle* or *charbon brillant*), which is dominated by vitrinite group macerals and appears glassy to the unaided eye, (2) clarain (*Glanzstreifenkohle* or *charbon semi-brillant*), which is composed of both vitrinite and exinite and has an appearance between that of vitrain and durain, (3) durain (*Mattkohle* or *charbon mat*), which is generally composed of fine-grained inertinites and exinites and has a dull, matlike lustre, and (4) fusain (*Faserkohle* or *charbon fibreux*), which is composed mainly of fusinite and resembles wood charcoal (because it soils the hands just as charcoal would).

Banded and nonbanded coals. The term coal type is also employed to distinguish between banded coals and nonbanded coals. Banded coals contain varying amounts of vitrinite and opaque material. They include bright coal, which contains more than 80 percent vitrinite, and splint coal, which contains more than 30 percent opaque matter. The nonbanded varieties include boghead coal, which has a high percentage of algal remains, and cannel coal with a high percentage of spores. The usage of all the above terms is quite subjective.

Ranking by coalification. *Hydrocarbon content.* The oldest coal-classification system was based on criteria of chemical composition. Developed in 1837 by the French chemist Henri-Victor Regnault, it was improved in later systems that classified coals on the basis of their hydrogen and carbon content. However, because the relationships between chemistry and other coal properties are complex, such classifications are rarely used for practical purposes today.

Chemical content and properties. The most commonly employed systems of classification are those based on analyses that can be performed relatively easily in the laboratory, as, for example, determining the percentage of volatile matter lost upon heating to about 950° C or the amount of heat released during combustion of the coal under standard conditions. Table 10 lists the ranks

Table 10: Classification of Coals by the American Society for Testing and Materials

					caloric value (moist, mineral-matter-free basis)*				
	fixed carbon percentage (dry, mineral-matter-free basis)		volatile matter percentage (dry, mineral-matter-free basis)		British thermal units per pound		megajoules per kilogram		
rank and group	equal to or greater than	less than	greater than	equal to or less than	equal to or greater than	less than	equal to or greater than	less than	agglomerating character
Anthracitic									
Meta-anthracite	98	. . .	. . .	2	. . .	. . .	. . .	. . .	nonagglomerating
Anthracite	92	98	2	8	. . .	. . .	. . .	. . .	
Semianthracite†	86	92	8	14	. . .	. . .	. . .	. . .	
Bituminous									
Low-volatile bituminous	78	86	14	22	. . .	. . .	. . .	. . .	commonly agglomerating§
Medium-volatile bituminous	69	78	22	31	. . .	. . .	. . .	. . .	
High-volatile A bituminous	. . .	69	31	. . .	14,000‡	. . .	32.6	. . .	
High-volatile B bituminous	. . .	. . .	. . .	. . .	13,000‡	14,000	30.2	32.6	
High-volatile C bituminous	. . .	. . .	. . .	. . .	11,500	13,000	26.7	30.2	
					10,500	11,500	24.4	26.7	agglomerating
Subbituminous									
Subbituminous A	. . .	. . .	. . .	. . .	10,500	11,500	24.4	26.7	nonagglomerating
Subbituminous B	. . .	. . .	. . .	. . .	9,500	10,500	22.1	24.4	
Subbituminous C	. . .	. . .	. . .	. . .	8,300	9,500	19.3	22.1	
Lignitic									
Lignite A	. . .	. . .	. . .	. . .	6,300	8,300	14.7	19.3	
Lignite B	. . .	. . .	. . .	. . .	. . .	6,300	. . .	14.7	

*Moist coal contains natural inherent moisture but does not include visible water on the surface. †If agglomerating, classify in low-volatile group of the bituminous rank. ‡Coals having 69 percent or more fixed carbon on the dry, mineral-matter-free basis are classified by fixed carbon, regardless of calorific value. §There may be nonagglomerating varieties in these groups of the bituminous rank, there are also notable exceptions in the high-volatile C bituminous group.

Source: *1993 Annual Book of ASTM Standards*, section 5, volume 5.05.

The ASTM system

assigned to coals by the American Society for Testing and Materials (ASTM) on the basis of fixed carbon content, volatile matter content, and calorific value. In addition to the major ranks (lignite, subbituminous, bituminous, and anthracite), each rank may be subdivided into coal groups, such as high-volatile A bituminous coal. Other designations, such as coking coal and steam coal, have been applied to coals, but they tend to differ from country to country.

Coal analyses may be presented in the form of "proximate" (a term derived from the word "approximate") and "ultimate" analyses, whose analytic conditions are prescribed by organizations such as the ASTM. A typical proximate analysis includes the moisture, ash, volatile matter, and fixed carbon contents. (Fixed carbon is the material, other than ash, that does not vaporize when heated in the absence of air. It is usually determined by subtracting the sum of the first three values—moisture, ash, and volatile matter—in weight percent from 100 percent.) It is important for economic reasons to know the moisture and ash contents of a coal because they do not contribute to the heating value of a coal. In most cases ash becomes an undesirable residue and a source of pollution, but for some purposes (*e.g.*, use as a chemical feedstock or for liquefaction) the presence of mineral matter may be desirable. Most of the heat value of a coal comes from its volatile matter, excluding moisture, and fixed carbon content. For most coals, it is necessary to measure the actual amount of heat released upon combustion (expressed in British thermal units [BTUs] per pound or megajoules per kilogram).

Ultimate analyses are used to determine the carbon, hydrogen, sulfur, nitrogen, ash, oxygen, and moisture contents of a coal. For specific applications, other chemical analyses may be employed. These may involve, for example, identifying the forms of sulfur present; sulfur may occur in the form of sulfide minerals (pyrite and marcasite), sulfate minerals (gypsum), or organically bound sulfur. In other cases the analyses may involve determining the trace elements present (*e.g.*, mercury, chlorine), which may influence the suitability of a coal for a particular purpose or help to establish methods for reducing environmental pollution and so forth.

Virtually all classification systems use the percentage of volatile matter present to distinguish coal ranks. In the ASTM classification (Table 10), high-volatile A bituminous (and higher ranks) are classified on the basis of their volatile matter content. Coals of lower rank are classified primarily on the basis of their heat values, because of their wide ranges in volatile matter content (including moisture). The so-called agglomerating character of a coal refers to its ability to soften and swell when heated and to form cokelike masses that are used in the manufacture of steel. The most suitable coals for agglomerating purposes are in the bituminous rank.

STRUCTURE AND PROPERTIES

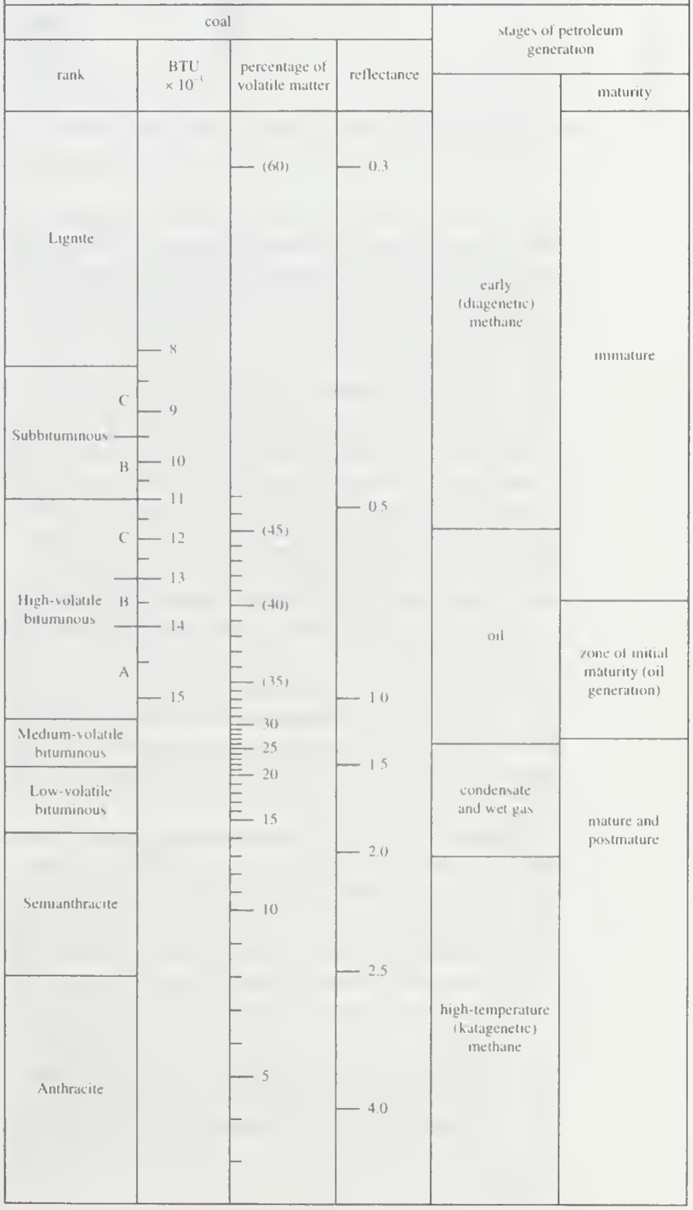
Volatile matter and fixed carbon

Organic compounds. The plant material from which coal is derived is composed of a complex mixture of organic compounds, including cellulose, lignin, fats, waxes, and tannins. As peat formation and coalification proceed, these compounds, which have more or less open structures, are broken down, and new compounds—primarily aromatic and hydroaromatic—are produced. These compounds are connected by cross-linking oxygen, sulfur, and molecules such as methylene. Some researchers refer to this complex array of cross-linked molecules as a "coal molecule." During coalification, volatile phases rich in hydrogen and oxygen (*e.g.*, water, carbon dioxide, and methane) are produced and escape from the mass; hence, the coal becomes progressively richer in carbon. The classification of coal by rank is based on these changes—*i.e.*, as coalification proceeds, the amount of volatile matter gradually decreases and the amount of fixed carbon increases. As volatiles are expelled, more carbon-to-carbon linkages occur in the remaining coal until, having reached the anthracite rank, it takes on many of the characteristics of the end product of the metamorphism of carbonaceous material—namely, graphite. Coals pass through several structural states, including a glassy (or "liquid") state, as the bonds between the aromatic (benzenelike compounds) nuclei increase.

Properties. Many of the properties of coal are strongly rank-dependent, although other factors such as maceral composition and the presence of mineral matter also influence its properties. Several techniques have been developed for studying the physical and chemical properties of coal, including density measurements, X-ray diffraction, scanning and transmission electron microscopy, infrared spectrophotometry, mass spectroscopy, gas chromatography, thermal analysis, and electrical, optical, and magnetic measurements.

Density. Knowledge of the physical properties of coal is important in coal preparation and utilization. For example, coal density ranges from approximately 1.1 to about 1.5 megagrams per cubic metre, or grams per cubic centimetre (1 megagram per cubic metre equals 1 gram per cubic centimetre). Coal is slightly denser than water (1.0 megagram per cubic metre) and significantly less dense than most rock and mineral matter (*e.g.*, shale has a density of about 2.7 megagrams per cubic metre and pyrite of 5.0 megagrams per cubic metre). Density differences make it possible to improve the quality of a coal by removing

Table 11: Relationship Between Coalification and the Organic Metamorphic Stages of Petroleum Generation



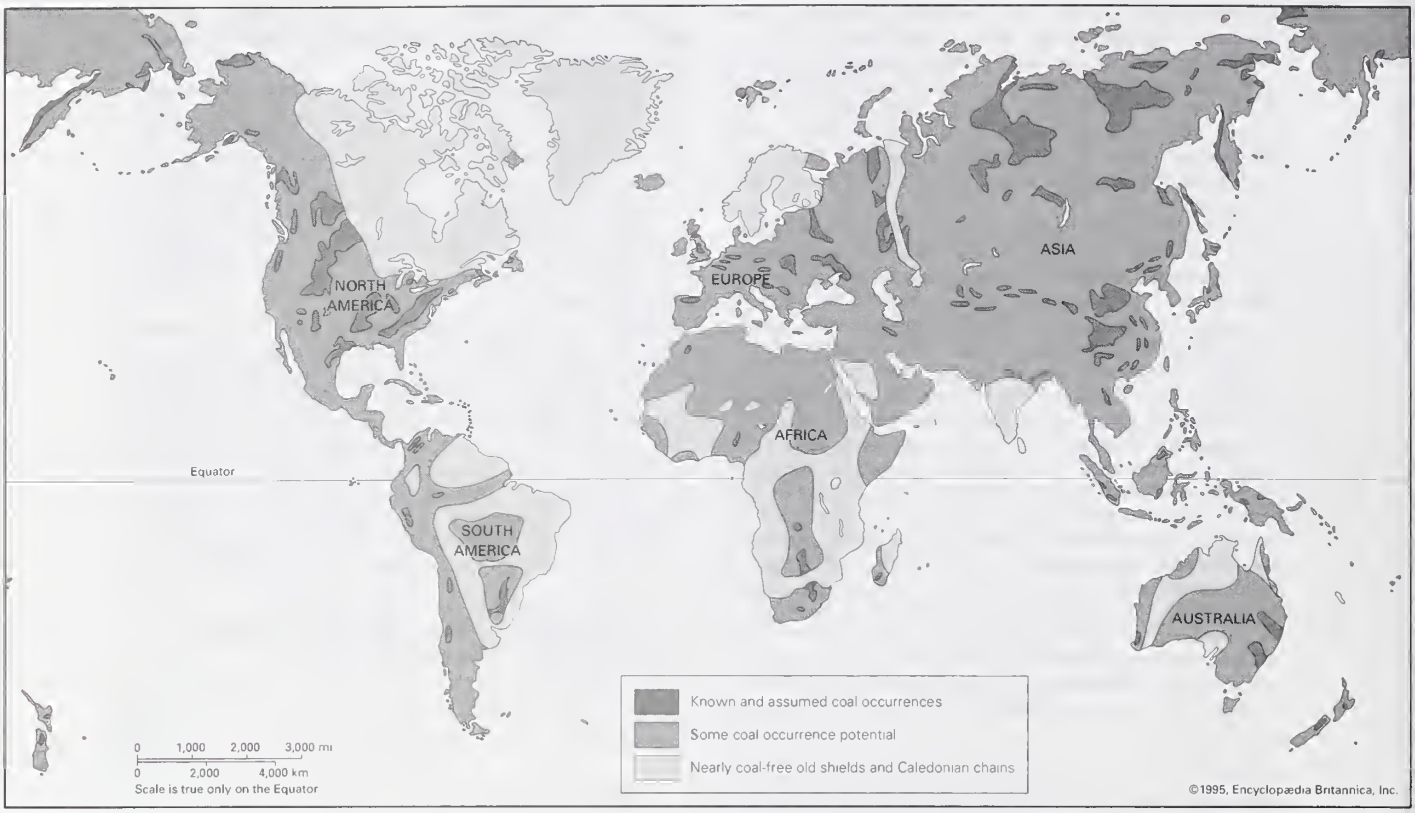

NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
AUSTRALIA
Equator
■ Known and assumed coal occurrences
▨ Some coal occurrence potential
□ Nearly coal-free old shields and Caledonian chains
0 1,000 2,000 3,000 mi
0 2,000 4,000 km
Scale is true only on the Equator
©1995, Encyclopædia Britannica, Inc.

Figure 12: Location of the most important coal occurrences on Earth.
Adapted from Gunter B. Fettweis, *World Coal Resources*, Elsevier Scientific Publishing Company, 1979

most of the rock matter and sulfide-rich fragments by means of heavy liquid separation (fragments with densities greater than about 1.5 megagrams per cubic metre settle out while the coal floats on top of the liquid) and devices such as cyclones and shaker tables, which also separate coal particles from rock and pyrite on the basis of their different densities.

Porosity. Coal density is controlled in part by the presence of pores that persist throughout coalification. Measurement of pore sizes and pore distribution is difficult; however, there appear to be three size ranges of pores: (1) macropores (diameter greater than 50 nanometres), (2) mesopores (diameter 2 to 50 nanometres), and (3) micropores (diameter less than 2 nanometres). (One nanometre is equal to 10^{-9} metre.) Most of the effective surface area of a coal—about 200 square metres per gram—is not on the outer surface of a piece of coal but is located inside the coal in its pores. The presence of pore space is important in the production of coke, gasification, liquefaction, and the generation of high-surface-area carbon for purifying water and gases. From the standpoint of safety, coal pores may contain significant amounts of adsorbed methane that may be released during mining operations and form explosive mixtures with air. The risk of explosion can by reduced by adequate ventilation during mining or by prior removal of coal-bed methane.

Reflectivity. One of the most important property of coal is its reflectivity (or reflectance)—*i.e.*, its ability to reflect light. Reflectivity is measured by shining a beam of monochromatic light (with a wavelength of 546 nanometres) on a polished surface of the vitrinite macerals in a coal sample and measuring the percentage of the light reflected with a photometer. Vitrinite is used because its reflectivity changes gradually with increasing rank. Fusinite reflectivities are too high due to its origin as charcoal, and exinites tend to disappear with increasing rank. Although little of the incident light is reflected (ranging from a few tenths of a percent to 10 to 12 percent), the value increases with rank and can be used to determine the rank of most coals indirectly (*i.e.*, without measuring the percentage of volatile matter). Typical reflectivity values are shown in Table 11.

The study of coals (and coaly particles called phyterals) in sedimentary basins containing oil and/or gas reveals a close relationship between coalification and the maturation of liquid and gaseous hydrocarbons (Table 11). During the initial stages of coalification (to a reflectivity of almost 0.5 and near the boundary between subbituminous and high-volatile C bituminous coal), hydrocarbon generation produces chiefly methane. The maximum generation of liquid petroleum occurs during the development of high-volatile bituminous coals (in the reflectivity range from roughly 0.5 to about 1.3). With increasing depth and temperature, petroleum liquids break down and, finally, only natural gas (methane) remains. Geologists can use coal reflectivity to anticipate the potential for finding liquid or gaseous hydrocarbons as they explore for petroleum.

Relationship between coalification and petroleum formation

Other properties. Other properties, such as hardness, grindability, ash-fusion temperature, and free-swelling index (a visual measurement of the amount of swelling that occurs when a coal sample is heated in a covered crucible), may affect coal mining and preparation, as well as the way in which a coal is utilized. Hardness and grindability determine the kinds of equipment used for mining, crushing, and grinding coals in addition to the amount of power consumed in their operation. Ash-fusion temperature influences furnace design and operating conditions. The free-swelling index provides preliminary information concerning the suitability of a coal for coke production.

WORLD DISTRIBUTION OF COAL

General occurrence. Coal is a widespread resource of energy and chemicals. Although terrestrial plants necessary for the development of coal did not become abundant until Carboniferous time, large sedimentary basins containing Carboniferous and younger rocks are known on virtually every continent (see Figure 12), including Antarctica (not shown on the map). The presence of coal deposits in regions that now have arctic or subarctic climates (such as Alaska and Siberia) is due to climatic changes and to the tectonic motion of crustal plates that moved ancient continental masses over the Earth's surface, sometimes through subtropical and even tropical regions. The absence of coal in the unshaded map areas (such as Greenland and

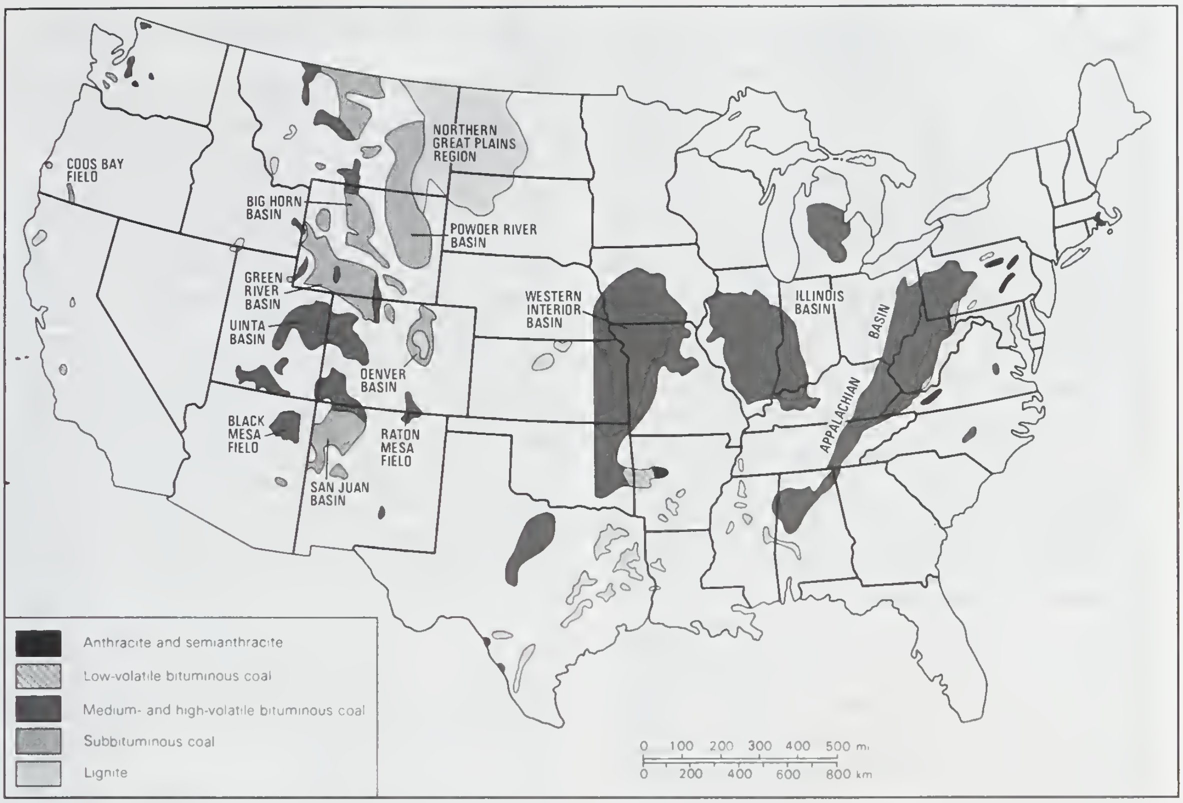


Figure 13: Coal-bearing areas of the conterminous United States.

Adapted from "Coal-Bearing Areas of the United States," *Coal Production 1985*, U.S. Energy Information Administration, 1985 and "Organic Fuel Deposits," *The National Atlas of the United States of America*, U.S. Department of Interior, 1970

much of northern Canada) results from the fact that the rocks found there are older than Carboniferous age and that these regions, known as shields, lacked the abundant terrestrial plant life needed for the formation of major coal deposits.

Resources and reserves. World coal resources are difficult to assess. Although some of the difficulty stems from the lack of accurate data on coal reserves and production figures for individual countries, two fundamental problems make estimates of worldwide reserves and production difficult and subjective.

Difficulty of defining reserves

The first problem concerns the definition of the term reserves. Ideally the reserves for any commodity, including coal, should provide a reasonably accurate estimate of the total recoverable amount. To be economically minable, a coal bed must have some minimum thickness (about 0.6 metre) and be buried less than some maximum depth (roughly 2,000 metres) below the Earth's surface. These values of thickness and depth are not fixed but change with the coal quality, demand, the ease with which overlying rocks can be removed (in surface mining) or a shaft sunk to reach the coal seam (in underground mining), and so forth. The development of new mining techniques can increase the amount of coal extracted relative to the amount that cannot be removed. For example, in underground mining (which accounts for about 60 percent of world coal production), conventional mining methods leave behind large pillars of coal to support the overlying rocks and may recover only about 50 percent of the coal present. On the other hand, longwall mining, in which the equipment removes continuous parallel bands of coal, may recover nearly all of the coal present.

Table 12: World Coal Resources and Reserves, by Major Coal-Producing Countries*

country	geologic resources	technically and economically recoverable reserves
Australia†	600,000	32,800
Canada†	323,036	4,242
China†	1,438,045	98,883
Germany†	246,800	34,419
India†	81,019	12,427
Poland†	139,750	59,600
South Africa	72,000	43,000
United Kingdom†	190,000	45,000
United States†	2,570,398	166,950
Russia, Ukraine, and Kazakhstan	4,860,000	109,900
Other countries	229,164	55,711
Total world	10,750,212	662,932

*In mtce, a unit of measure equaling 1,000,000 tons of coal equivalent.
†Participants of the World Coal Study (WOCOL).
Sources: World Energy Conference, WOCOL Country Reports, and Simon Walker, *Major Coalfields of the World* (1993).

The second problem concerning the estimation of reserves is the rate at which a commodity is consumed. When considering coal as a resource, the number of years that coal will be available may be more important than the total amount of coal reserves. Estimates of how long coal reserves will last range from a few hundred years to more than 1,000 years, depending on technological developments and assumed rates of consumption.

Table 12 indicates the extent of world coal resources using a simple twofold division: (1) geologic resources, including all coal resources whether recoverable or not, and (2) technically and economically recoverable reserves. Note that the latter category constitutes less than 10 percent of the estimated total world coal. The amounts in the table are given in million tons of coal equivalent. One ton of coal equivalent equals 1 metric ton (2,205 pounds) of coal with a heating value of 29.3 megajoules per kilogram (12,600 BTUs per pound). The estimates in Table 12 suggest that the United States has the largest amount of recoverable coal. (The coalfields of the conterminous United States are shown in Figure 13.) Nearly 60 percent of the world's recoverable coal resources are controlled by the United States, the former Soviet republics of Russia, Ukraine, and Kazakhstan, and China. (O.C.K.)

BIBLIOGRAPHY

Oil and natural gas. ROBERT H. DOTT, SR., and MERRILL J. REYNOLDS (comps.), *Sourcebook for Petroleum Geology* (1969), is an excellent collection of petroleum-geology scientific concepts. EDGAR WESLEY OWEN, *Trek of the Oil Finders: A History of Exploration for Petroleum* (1975); and HAROLD F. WILLIAMSON and ARNOLD R. DAUM, *The American Petroleum Industry*, 2 vol. (1959–63, reprinted 1981), relate the history

of petroleum geology and the development of the oil industry. KENNETH K. LANDES, *Petroleum Geology,* 2nd ed. (1959, reprinted 1975); L.W. LEROY, D.O. LEROY, and J.W. RAESE (eds.), *Subsurface Geology: Petroleum, Mining, Construction,* 4th ed. (1977); B.P. TISSOT and D.H. WELTE, *Petroleum Formation and Occurrence,* 2nd rev. ed. (1984); and JOHN M. HUNT, *Petroleum Geochemistry and Geology* (1979), are sources of information on theories of the origin and accumulation of petroleum, as well as on the practical applications of scientific knowledge to petroleum problems. E.N. TIRATSOO, *Oilfields of the World,* 3rd ed., rev. (1986), describes the nature and distribution of the major oil fields of the world. C.D. MASTERS, D.H. ROOT, and E.D. ATTANASI, "Resource Constraints in Petroleum Production Potential," *Science,* 253(5016):146–152 (July 12, 1991), contains U.S. Geological Survey assessments of world oil and gas resources. JOSEPH P. RIVA, JR., *Exploration Opportunities in Latin America* (1992), provides information on Latin American oil and gas reserves, resources, and production potential, and *Petroleum Exploration Opportunities in the Former Soviet Union* (1994), contains similar information for each of the former Soviet republics. MICHEL T. HALBOUTY (ed.), *Future Petroleum Provinces of the World* (1986), contains an annotated list of the world's giant oil and gas fields. KENNETH K. LANDES, *Petroleum Geology of the United States* (1970), describes in detail petroleum geology on a state-by-state basis. JOSEPH P. RIVA, JR., JOHN J. SCHANZ, JR., and JOHN G. ELLIS, *U.S. Conventional Oil and Gas Production: Prospects to the Year 2000* (1985), assesses U.S. domestic reserves, resources, and production potential. JOSEPH P. RIVA, JR., *World Petroleum Resources and Reserves* (1983), treats the formation and accumulation of conventional and unconventional petroleum, exploration and production methods, and petroleum basin geology. G.D. HOBSON (ed.), *Modern Petroleum Technology,* 5th ed., 2 vol. (1984), discusses petroleum engineering, including production, refining, and transport. FILLMORE C.F. EARNEY, *Petroleum and Hard Minerals from the Sea* (1980), examines worldwide development of seabed resources with emphasis on offshore petroleum.

JAMES A. CLARK, *The Chronological History of the Petroleum and Natural Gas Industries* (1963), gives a detailed chronology of both technical and human facts. MALCOLM W.H. PEEBLES, *Evolution of the Gas Industry* (1980), provides international, historical, and technological developments. ARLON R. TUSSING and CONNIE C. BARLOW, *The Natural Gas Industry: Evolution, Structure, and Economics* (1984), examines the natural gas industry in the United States. E.L. RAWLINS and M.A. SCHELLHARDT, *Back-Pressure Data on Natural-Gas Wells and Their Application to Production Practices* (1935, reissued 1970), is a classic report of the U.S. Bureau of Mines describing and explaining the back-pressure method, with an analysis of data for more than 500 gas wells. MORRIS MUSKAT, *Physical Principles of Oil Production,* 2nd ed. (1981), and *The Flow of Homogeneous Fluids Through Porous Media* (1937, reprinted 1982), are fundamental works on the basic principles of gas and petroleum.

Collections of scientific papers may be found in G.D. HOBSON (ed.), *Developments in Petroleum Geology,* 2 vol. (1977–80); and in publications of the American Association of Petroleum Geologists, including the *AAPG Bulletin* (monthly), the October issue of which contains an annual review of significant exploration and production activity; and the *AAPG Memoir* (irregular). *Basic Petroleum Data Book* (three per year); and *Minerals Yearbook,* prepared by the U.S. Bureau of Mines, include annual statistical reviews of the petroleum industry. Each year maps, production figures, and geologic data are published in August by *World Oil* and in December by the *Oil and Gas Journal.* (J.P.Ri./G.I.A.)

Heavy oils, tar sands, and oil shales. Current information concerning the characteristic properties of oil shales, their sources, and special problems of exploitation may be found in the proceedings of meetings, such as *Oil Shale Symposium Proceedings* (annual); material originating at symposia chaired by PAUL B. TARMAN, *Synthetic Fuels from Oil Shale* (1980), *Synthetic Fuels from Oil Shale II* (1982), and *Synthetic Fuels from Oil Shale and Tar Sands* (1983); and H.C. STAUFFER (ed.), *Oil Shale, Tar Sands, and Related Materials* (1981), symposium papers, including essays on oil shale cracking and retorting. General works include KEN P. CHONG and JOHN WARD SMITH (eds.), *Mechanics of Oil Shale* (1984), a collection of summary papers on the exploitation of oil shales; T.F. YEN and GEORGE V. CHILINGARIAN (eds.), *Oil Shale* (1976), background essays on different aspects of oil shale technology and science; PAUL L. RUSSELL, *History of Western Oil Shale* (1980); and PERRY NOWACKI (ed.), *Oil Shale Technical Data Handbook* (1981). The transitional character of kerogen rocks and their limnological and stratigraphical properties are treated in BERNARD DURAND (ed.), *Kerogen: Insoluble Organic Matter from Sedimentary Rocks* (1980); BARTHOLOMEW NAGY and UMBERTO COLOMBO (eds.), *Fundamental Aspects of Petroleum Geochemistry* (1967); and A.I. LEVORSEN, *Geology of Petroleum,* 2nd rev. ed. (1967).

Data on world distribution, exploitation, and technology are included in FERDINAND MAYER, *Weltatlas Erdöl und Erdgas,* 2nd ed. (1976); and T.F. YEN (ed.), *Science and Technology of Oil Shale* (1976). RICHARD F. MEYER (ed.), *Exploration for Heavy Crude Oil and Natural Bitumen* (1987), contains information on the size and distribution of the world's largest accumulations of heavy crude and bitumen. Information on oil shales, tar sands, and heavy oils is presented in *Zeitschrift für Angewandte Geologie* (monthly); *Erdöl-Erdgas* (monthly); and *Oil and Gas Journal* (weekly). (J.P.Ri.)

Coals. HOWARD N. EAVENSON, *Coal Through the Ages,* 2nd ed., rev. (1942), provides information about the early history of coal mining and coal utilization. MARIE C. STOPES and R.V. WHEELER, *Monograph on the Constitution of Coal* (1918), is a classic work on the organic composition of coals. Information on this and other subjects treated in the present article and numerous references to earlier literature may be found in D.W. VAN KREVELEN, *Coal: Typology, Physics, Chemistry, Constitution,* 3rd, completely rev. ed. (1993); E. STACH *et al., Stach's Textbook of Coal Petrology,* 3rd rev. and enlarged ed. (1982; originally published in German, 1935); and DUNCAN MURCHISON and T. STANLEY WESTOLL (eds.), *Coal and Coal-Bearing Strata* (1968). ROBERT A. MEYERS (ed.), *Coal Structure* (1982), examines the nature and origin of coal structure and porosity. SIMON WALKER, *Major Coalfields of the World* (1993), provides information about coal resources worldwide, including in the Commonwealth of Independent States. Methods for estimating world coal reserves, the present status of coal reserves, and the prospects for future development of coal resources are examined in GÜNTER B. FETTWEIS, *World Coal Resources: Methods of Assessment and Resources* (1979; originally published in German, 1976); and CARROLL L. WILSON, *Coal—Bridge to the Future: A Report of the World Coal Study* (1980). DOUGLAS C. PETERS (ed.), *Geology in Coal Resource Utilization* (1991), provides information concerning coal resources, reserve estimation, coal utilization, and the environment. (O.C.K.)

Fungi

The kingdom Fungi (Mycota) comprises a large group of eukaryotic organisms having two common characteristics: anatomically, their principal mode of vegetative growth is through mycelium; physiologically, their nutrition is based on absorption of organic matter. Although historically included in the plant kingdom, fungi lack chlorophyll and other structures common among true plants and have therefore been placed in a separate kingdom. Fungi are the culmination of a major direction in evolution distinctly different from that of plants or animals; this evolutionary line was established by organisms whose nutrition was based on absorption of organic matter. Fungi are among the most widely distributed organisms on Earth and are of great importance. The fungi include yeasts, rusts, smuts, mildews, molds, mushrooms, and others. Many fungi are free-living in soil or water; others form parasitic or symbiotic relationships with plants or animals, respectively. (Slime molds, straddling the animal and plant worlds, are treated in the article PROTISTS.)

The mushrooms, by no means the most numerous or economically significant of the fungi, are the most conspicuous members of the group; thus, the Latin word for mushroom, *fungus* (plural *fungi*), has come to stand for the whole group. Similarly, the study of fungi is known as mycology—a broad application of the Greek word for mushroom, *mykēs*. Fungi other than mushrooms are sometimes collectively called molds, although this term is better restricted to fungi of the sort represented by bread mold.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 313, and the *Index*.

This article is divided into the following sections:

GENERAL FEATURES

Basic structure

A typical fungus consists of a mass of branched, tubular filaments enclosed by a rigid cell wall. The filaments, called hyphae (singular hypha), branch repeatedly into a complicated, radially-expanding network called the mycelium, which makes up the thallus, or undifferentiated body, of the typical fungus. Some fungi, notably the yeasts, do not form a mycelium but grow as individual cells that multiply by budding or, in certain species, by fission. The mycelium grows by utilizing nutrients from the environment and, upon reaching a certain stage of maturity, forms—either directly or in special fruiting bodies—reproductive cells called spores. The spores are released and dispersed by a wide variety of passive or active mechanisms; upon reaching a suitable substrate, the spores germinate and develop hyphae that grow, branch repeatedly, and become the mycelium of the new individual. Fungal growth is mainly confined to the tips of the hyphae.

All fungal structures are, therefore, made of hyphae or portions of hyphae; for example, a mushroom—visibly consisting of a stem supporting a cap with gills on the underside—is constructed entirely of microscopic filaments intricately associated and interwoven. The mushroom is the fruiting body of the mycelium and bears the spores; the main body of the fungus is underground and consists of a huge network of hyphae—the mycelium—spread over a very large area, often several metres (yards) in diameter. This mycelium obtains food from organic matter in the soil and grows outward, just below the surface, in a circular fashion. In certain species the hyphal branches at the edge of the mycelium become organized at intervals into elaborate tissues that develop above ground into mushrooms. Such a circle of mushrooms is known as a fairy ring (see Figure 2) because, in the Middle Ages, it was believed to represent the path of dancing fairies. The ring marks the periphery of an enormous fungus colony, which, if undisturbed, continues to produce ever wider fairy rings year after year.

Size range and diversity of structure. The part of a fungus that is generally visible is the fruiting body, or sporophore. Sporophores vary greatly in size, shape, colour, and longevity. Some are microscopic and completely invisible to the unaided eye; others are no larger than a pin head; still others are gigantic structures. Among the largest sporophores are those of the mushrooms, bracket fungi, and puffballs. Some mushrooms reach a diameter of 20 to 25 centimetres (8 to 10 inches) and a height of 25 to 30 centimetres. Bracket fungi 40 centimetres or more in diameter are not uncommon, and puffballs often exceed that size. The largest puffballs on record measured 150 centimetres in diameter. The number of spores within such giants reaches several trillion.

Distribution and abundance. Fungi are either terrestrial or aquatic, the latter living in freshwater or marine environments. Freshwater species usually do not tolerate high degrees of salinity in nature, although some are found in slightly brackish water; most prefer clean, cool waters, but a few thrive in highly polluted streams. The soil furnishes an ideal habitat for a large number of species. Fungi are found in all temperate and tropical regions of the world where there is sufficient moisture to enable them to grow. They have also been reported from the Arctic and Antarctic regions but are much rarer there, being replaced by lichens, organisms that have fungal and algal components living in symbiosis (see below *Lichens*). In general, fungi are abundant in moist habitats where organic matter is plentiful and less abundant in drier areas or in habitats with little or no organic matter. About 50,000 species of fungi have been identified and described, but it has been estimated that there may be as many as 100,000 to 250,000 total species. Previously undescribed species are constantly being discovered and, as the tropics become more thoroughly explored, many more new species will undoubtedly be found.

Number of species

Importance. In 1928 a green mold accidentally grew in a culture dish of *Staphylococcus* bacteria that the bacteriologist Alexander Fleming was studying in a London hospital. The fungus colony that developed inhibited the growth of the bacteria. Such unavoidable contamination certainly had occurred many times before in laboratories through-

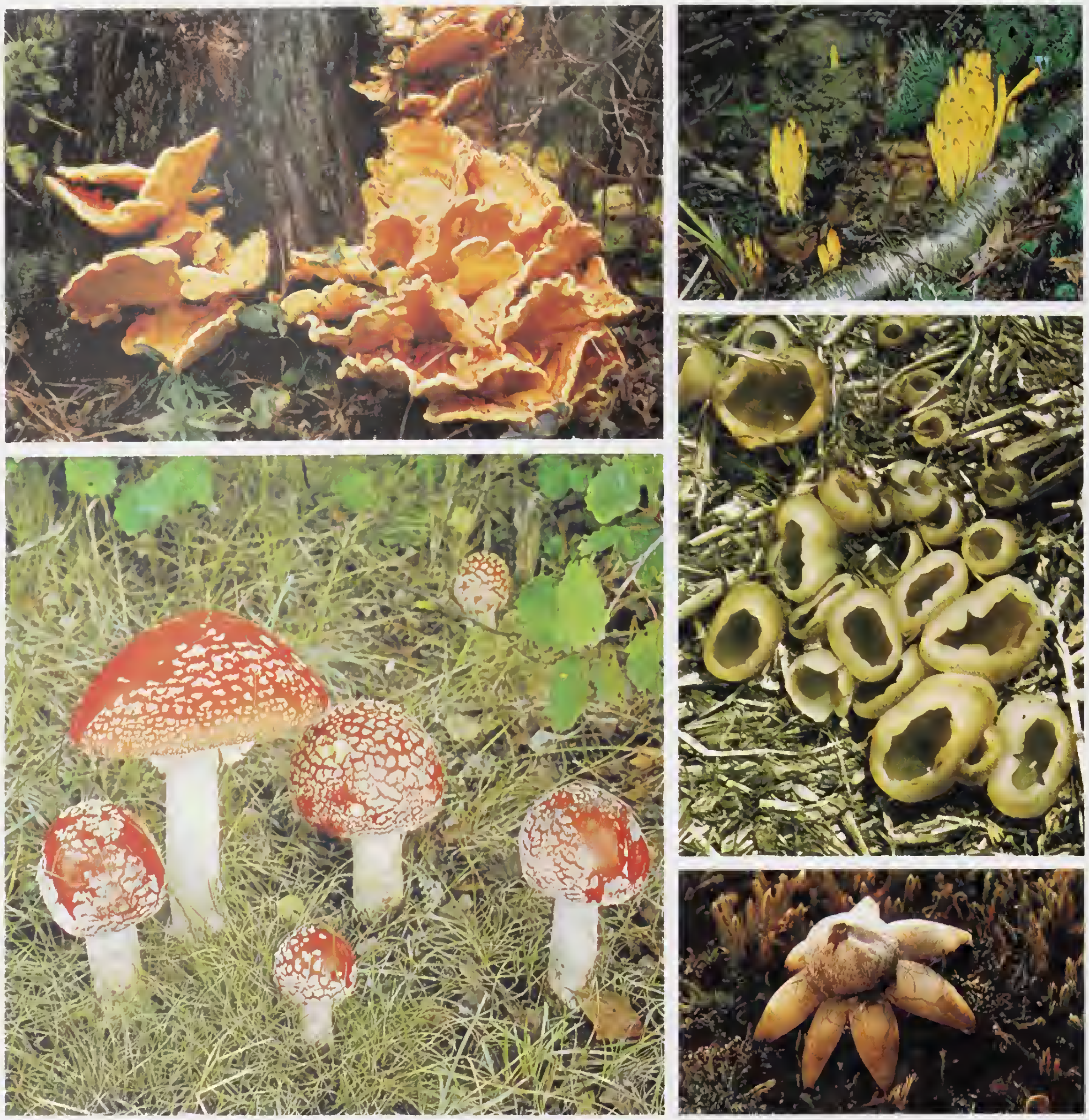


Figure 1: *Fruiting structures among saprophytic higher fungi.*
Plant bodies of these fungi consist of masses (mycelia) of threadlike filaments (hyphae) growing in soil or wood. All the fungi shown are Basidiomycetes except *Peziza*, an ascomycete. (Top left) Bracket fungus, chicken of the woods (*Polyporus sulphureus*), growing on wood. (Top right) Club fungus (*Clavaria*) growing in soil. (Centre right) Cup fungus (*Peziza vesiculosa*) on woody forest litter. (Bottom left) Gill fungus, the poisonous fly agaric (*Amanita muscaria*), growing on soil. (Bottom right) Puffball, the earthstar (*Geastrum*), growing on moist soil among mosses.

(Top left) H.S Knighton, (top right) Ken Brate—Photo Researchers, (centre right) H.R Allen—EB Inc , (bottom left) Sven Samelius, (bottom right) Larry West—The National Audubon Society Collection/Photo Researchers

out the world, but the people who may have seen such cultures probably regarded them as contaminated plates to be discarded as soon as possible. Fleming, however, carefully recorded his observation and in 1929 published a scientific report announcing the discovery of penicillin, the first of a series of antibiotics—many of them derived from fungi—that have revolutionized medical practice.

Ergotism

In 1951 a strange disease broke out in the small French village of Pont-Saint-Esprit, and several persons died. Doctors were baffled by the mysterious malady until it was recognized as a form of "St. Anthony's fire"—ergotism—that had resulted from eating bread made from contaminated flour. Ergotism was prevalent in northern Europe in the Middle Ages, particularly in regions of high rye-bread consumption; modern grain-cleaning and milling methods have practically eliminated the disease.

The cause of ergotism is ergot—a fungus. More precisely, ergot is a sclerotium (plural sclerotia), a special part of a fungus that develops on grasses and especially on rye. The wind carries the fungal spores to the flowers of the rye, where the spores germinate, infect and destroy the ovaries of the plant, and replace them with masses of microscopic threads cemented together into a hard fungal structure shaped like a rye kernel but considerably larger and darker. This is ergot, and it contains a number of poisonous organic compounds called alkaloids. A mature head of rye may carry several ergots in addition to noninfected kernels. When the grain is harvested, much of the ergot falls to the ground, but some remains on the plants and is mixed with the grain. If the ergot is not removed before milling, the ergotized flour would be converted into bread and other food products and consumed; St. Anthony's fire—for which no cure is known—is the result. The ergot that falls to the ground may be the source of more trouble. Cattle put to graze in the rye fields after harvest are likely to consume enough ergot to bring on abortion of fetuses or death. In the spring, when the rye is in bloom, the ergot remaining on the ground produces tiny, black, mushroomlike bodies that expel large numbers of spores to start a new series of infections.

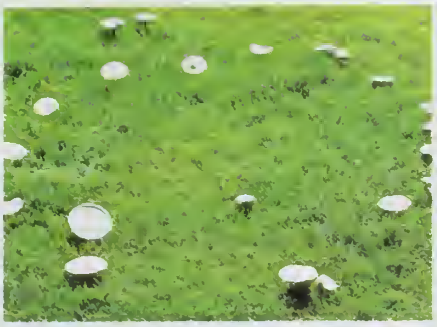


Figure 2: Fairy ring of mushrooms (*Chlorophyllum molybditis*).
Walter Dawn

Among the many interesting chemicals in ergot is lysergic acid, the active principle of the psychedelic drug lysergic acid diethylamide (LSD). Here, then, is a single fungus that can reduce crop yields, cause abortion in cattle, sicken and sometimes kill people, and be used as a source of LSD. On the credit side, ergot provides medical science with drugs useful in inducing labour in pregnant women and in controlling hemorrhage after birth.

The systematic study of fungi began 250 years ago, but humans have been indirectly aware of fungal activity since the first loaf of leavened bread was baked and the first tub of grape must was turned into wine. Yet, even now, few people realize that they are almost constantly either benefited or harmed by these organisms. Fungi are everywhere in very large numbers—in the soil and the air, in lakes, rivers, and seas, on and within plants and animals, in food and clothing, and in the human body; it is this that makes them so important in the human environment. Together with bacteria, fungi are responsible for the disintegration of organic matter and the release, into the soil or atmosphere, of the carbon, oxygen, nitrogen, and phosphorus that otherwise would be forever locked up in undecomposed organic matter. Fungi are essential to many household and industrial processes, notably the making of bread, wine, beer, and certain cheeses. They are used in the production of a number of organic acids, enzymes (biological catalysts), and vitamins and are the sources of a number of antibiotics besides penicillin. Fungi are also used as food: mushrooms, morels, and truffles are epicurean delicacies.

Universal presence of fungi

Studies of fungi have greatly contributed to the accumulation of fundamental knowledge in biology. Current knowledge of biochemistry and cellular metabolism was derived in part from studies of ordinary baker's or brewer's yeast (*Saccharomyces cerevisiae*). Some of these pioneering discoveries were made at the end of the 19th century and continued during the first half of the 20th. From 1920 through the 1940s, geneticists and biochemists who studied mutants of the red bread mold, *Neurospora*, established the one-gene–one-enzyme theory and laid the foundation of modern genetics. These and other fungi continue to be useful for studying cell and molecular biology, genetic engineering, and other basic disciplines of biology.

Although humans have benefited from the activities of certain fungi, they have also learned to look for villains among them. Numerous instances of destruction, disease, and death are directly traceable to fungi. The chestnut forests of the United States, for example, have been destroyed by the chestnut blight fungus, and the elms in both the United States and Europe have been devastated by the fungus that causes Dutch elm disease. Blights, rusts, wilts, and smuts destroy crops as land is cultivated more intensively, though increasingly effective control measures have been developed. Fungi cause various human diseases—*e.g.,* athlete's foot and ringworm, candidiasis and aspergillosis, histoplasmosis and coccidioidomycosis. Mildews, molds, and rots ruin clothing and foods.

Disease-causing fungi

Ancient peoples were familiar with the ravages of fungi in agriculture but attributed these diseases to the wrath of the gods. The Romans even designated a particular deity, Robigus, as the god of rust and, in an effort to propitiate him, organized an annual festival, the Robigalia, in his honour.

The mushrooms, because of their size, are easily seen in fields and forests and consequently were the only fungi known before the invention of the microscope in the 17th century. The microscope made it possible to recognize and identify the great variety of fungal species living on dead or live organic matter, the existence of which could not have been apprehended by the unaided eye.

NATURAL HISTORY

Following a period of intensive growth, which in most fungi consists in the development of an extensive mycelium, fungi enter their reproductive phase by forming and releasing vast quantities of spores. Spores are usually single, but sometimes multiple, cells produced either by fragmentation of the mycelium or within specialized structures (sporangia, gametangia, sporophores, etc.). The main types of reproductive structures also distinguish the major classes of fungi; finer details in these reproductive structures are commonly used to distinguish species among the genera. Spores may be produced either directly by asexual methods or indirectly through sexual reproduction. Sexual reproduction in fungi, as in other living organisms, involves the fusion of two nuclei that are brought together when two sex cells (gametes) unite. Asexual reproduction, which is simpler and more direct, may be accomplished by various methods.

Asexual reproduction. Typically in asexual reproduction, a single individual gives rise to a genetic duplicate of the progenitor without a genetic contribution from another individual. Perhaps the simplest method of reproduction of fungi is by fragmentation of the thallus, the body of a fungus. Some yeasts, which are single-celled fungi, reproduce by simple cell division, in which one cell undergoes nuclear division and splits into two daughter cells; after some growth, these cells divide, and, eventually, a population of cells forms. In filamentous fungi the mycelium may fragment into a number of segments, each of which is capable of growing into a new individual. In the laboratory, fungi are commonly propagated on a layer of solid nutrient agar inoculated either with spores or with fragments of mycelium.

Budding, which is another method of asexual reproduction, occurs in most yeasts and in some filamentous fungi. In this process, a bud develops on the surface of either the yeast cell or the hypha, its cytoplasm being continuous with that of the parent cell. The nucleus of the parent cell then divides; one of the daughter nuclei migrates into the bud, and the other remains in the parent cell. The latter is capable of producing many buds over its surface by continuous synthesis of cytoplasm and repeated nuclear divisions. After a bud develops to a certain point and even before it is severed from the parent cell, it is itself capable of budding by the same process. In this way, a chain of cells may be produced. Eventually, the individual buds pinch off the parent cell and become individual yeast cells. Buds that are pinched off a hypha of a filamentous fungus behave as spores; that is, they germinate, each giving rise to a structure called a germ tube, which develops into a new hypha.

Budding

Although fragmentation, fission, and budding are methods of asexual reproduction in a number of fungi, the majority reproduce asexually by the formation of spores. Spores that are produced asexually are often termed mitospores, and such spores are produced in a variety of ways. For a more detailed discussion of spores, see below *Form and function*.

Sexual reproduction. Sexual reproduction, an important source of genetic variability, allows the fungus to adapt to new environments. The process of sexual reproduction among the fungi is in many ways unique. Whereas nuclear division in other eukaryotes (animals, plants, and protists) involves the dissolution and reformation of the nuclear membrane, the nuclear membrane remains intact throughout the process in fungi, although gaps in its integrity are found in some species. The nucleus of the fungus becomes pinched at its midpoint, and the diploid chromosomes are pulled apart by spindle fibres formed within the intact nucleus. The nucleolus is usually also retained and divided between the daughter cells, although it may be expelled from the nucleus, or it may be dispersed within the nucleus but detectable.

Stages of sexual reproduction

Sexual reproduction in the fungi consists of three sequential stages: plasmogamy, karyogamy, and meiosis. The diploid chromosomes are pulled apart into two daughter cells, each containing a single set of chromosomes (a haploid state). Plasmogamy, the fusion of two protoplasts (the contents of the two cells), brings together two compatible haploid nuclei. At this point, two nuclear types are present in the same cell, a condition called dikaryotic, but the nuclei have not yet fused. Karyogamy results in the fusion of these haploid nuclei and the formation of a diploid nucleus (*i.e.,* a nucleus containing two sets of chromosomes, one from each parent). The cell formed by karyogamy is called the zygote. In most fungi, the zygote is the only cell in the entire life cycle that is diploid. The dikaryotic state that results from plasmogamy is often a prominent condition in fungi and may be prolonged over several generations. In the lower fungi, karyogamy usually follows plasmogamy almost immediately. In the more evolved fungi, however, karyogamy is separated from plasmogamy. Once karyogamy has occurred, meiosis (cell division that reduces the chromosome number to one set per cell) generally follows immediately and restores the haploid phase. Either the haploid nuclei that result from meiosis or their immediate progeny are generally incorporated in spores called meiospores.

Fungi employ a variety of methods to bring together two compatible haploid nuclei (plasmogamy). Some produce specialized sex cells (gametes) that are released from differentiated sex organs called gametangia. In other fungi two gametangia come in contact, and nuclei pass from the male gametangium into the female, thus assuming the function of gametes. In still other fungi the gametangia themselves may fuse in order to bring their nuclei together. Finally, some of the most advanced fungi produce no gametangia at all; the somatic (vegetative) hyphae take over the sexual function, come in contact, fuse, and exchange nuclei.

Fungi in which a single individual bears both male and female gametangia are hermaphroditic fungi. Rarely, gametangia of different sexes are produced by separate individuals, one a male, the other a female. Such species are termed dioecious. Dioecious species usually produce sex organs only in the presence of an individual of the opposite sex.

Mating types

Incompatibility. Many fungi produce differentiated male and female organs on the same thallus but do not undergo self-fertilization because their sex organs are incompatible. Such fungi require the presence of thalli of different mating types in order for sexual fusion to take place. The simplest form of this mechanism occurs in fungi in which there are two mating types, often designated + and − (or *A* and *a*). Gametes produced by one type of thallus are compatible only with gametes produced by the other type. Such fungi are said to be heterothallic. Many fungi, however, are homothallic; *i.e.,* sex organs produced by a single thallus are self-compatible, and a second thallus is unnecessary for sexual reproduction. Some of the most complex fungi (*e.g.,* mushrooms) do not develop differentiated sex organs; rather, the sexual function is carried out by their somatic hyphae, which unite and bring together compatible nuclei in preparation for fusion. Homothallism and heterothallism are encountered in these fungi, as well as in those in which sex organs are easily distinguishable. Compatibility, therefore, refers to a physiological differentiation, and sex refers to a morphological (structural) one; the two phenomena, although related, are not synonymous.

Sexual pheromones (hormones). The formation of sex organs in fungi is often induced by specific organic substances. Although called sex hormones when first discovered, these organic substances are actually sex pheromones, chemicals produced by one partner to elicit a sexual response in the other. In *Allomyces* (Chytridiomycetes) a pheromone named sirenin, secreted by the female gametes, attracts the male gametes, which swim toward the former and fuse with them. In *Achlya* (Oomycetes) a sterol pheromone called antheridiol induces the formation of gametangia and attracts the male to the female. In the Zygomycetes, in which the gametangia are usually not differentiated structurally, a complex biochemical interplay between mating types produces trisporic acid, a pheromone that induces the formation of specialized aerial hyphae. An unidentified volatile factor causes the tips of opposite mating aerial hyphae to grow toward each other and fuse. In yeasts (Ascomycetes and Basidiomycetes) the pheromones are small peptides.

Life cycle. In the life cycle of a sexually reproducing fungus, a haploid phase alternates with a diploid phase (see, for example, Figure 3). The haploid phase ends with nuclear fusion, and the diploid phase begins with the formation of the zygote (the diploid cell resulting from fusion of two haploid sex cells). A special kind of nuclear division, termed meiosis or reduction division, restores the haploid number of chromosomes and initiates the haploid phase, which produces the gametes. In the majority of fungi, all structures are haploid except the zygote. Nuclear fusion takes place at the time of zygote formation, and meiosis follows immediately. Only in the water mold *Allomyces* and a few related genera and in some yeasts is alternation of a haploid thallus with a diploid thallus definitely known. In the class Oomycetes (Figure 4) the thallus is diploid, and meiosis takes place just before the formation of the gametes.

From C J Alexopoulos, *Introductory Mycology*, 2nd ed , copyright © (1962) by John Wiley and Sons, Inc , reprinted by permission

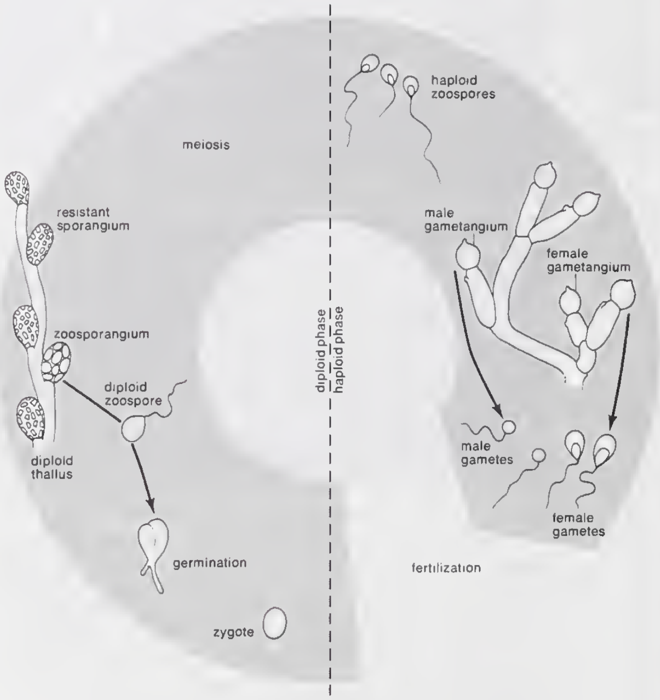


Figure 3: Life cycle of the water mold *Allomyces macrogynus*, a chytridiomycete.

In the higher fungi a third condition is interspersed between the haploid and diploid phases of the life cycle. In these fungi, plasmogamy (fusion of the cellular contents of two hyphae but not of the two haploid nuclei) results in dikaryotic hyphae in which each cell contains two haploid nuclei, one from each parent. Eventually, the nuclear pair fuses to form the diploid nucleus and thus the zygote. In the Basidiomycetes, the most advanced class of fungi, the binucleate cells divide successively and give rise to a binucleate mycelium, which is the main assimilative phase of the life cycle. It is the binucleate mycelium that eventually forms the basidia—the stalked fruiting bodies in which nuclear fusion and meiosis take place prior to the formation of the basidiospores.

Fungi usually reproduce both sexually and asexually. The asexual cycle produces mitospores, and the sexual cycle produces meiospores. Even though both types of spores are

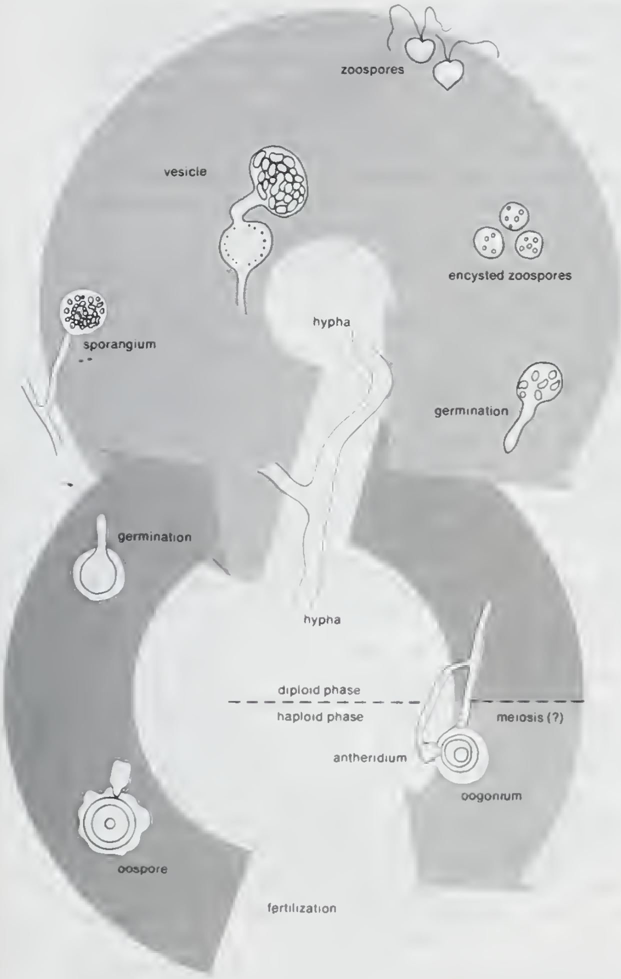


Figure 4: Life cycle of *Pythium debaryanum*, an oomycete.
From C J Alexopoulos, *Introductory Mycology* 2nd ed copyright © 1962 by John Wiley & Sons, Inc reprinted by permission

produced by the same mycelium, they are very different in form and easily distinguished (see below *Sporophores and spores*). The asexual phase usually precedes the sexual phase in the life cycle and may be repeated frequently before the sexual phase appears.

Some fungi differ from others in their lack of one or the other of these stages; some reproduce only sexually (except for fragmentation, which is common in most fungi), and many only asexually. A number of fungi exhibit the phenomenon of parasexuality, in which processes comparable to plasmogamy, karyogamy, and meiosis take place but not at a specified time or at specified points in the life cycle of the organism. Parasexuality is characterized by the prevalence of heterokaryosis in a mycelium—*i.e.*, the presence, side by side, of nuclei of different genetic composition.

Parasexuality

Ecology. Relatively little is known of the effects of the environment on the distribution of fungi that utilize dead organic material as food (saprobic fungi; see below *Nutrition*). The availability of organic food is certainly one of the factors controlling such distribution. A great number of fungi appear able to utilize most types of organic materials, such as lignin, cellulose, or other polysaccharides, which have been added to soils or waters by dead vegetation. It follows that most saprobic fungi may be expected to be cosmopolitan, at least in habitats with a sufficient organic content to support fungal growth. Whereas a great many saprobes are cosmopolitan, some are strictly tropical and others are strictly temperate-zone forms. Fungi with specific nutritional requirements are, of course, more localized.

Moisture and temperature are two additional ecological factors that are important in determining the distribution of fungi. Laboratory studies have shown that many, perhaps the majority, of fungi are mesophilic; *i.e.*, they have an optimum growth temperature of 20°–30° C (68°–86° F). Thermophilic species are able to grow at 50° C (122° F) or higher but are unable to grow below 30° C. Although the optimum temperature for growth of most fungi lies at or above 20° C (68° F), a large number of species are able to grow close to or below 0° C (32° F). The so-called snow molds and the fungi that cause spoilage of refrigerated foods are examples of this group. Obviously, temperature relationships influence the distribution of various species. Certain other effects of temperature are also important factors in determining the habitats of fungi. Many coprophilous (dung-inhabiting) fungi, for example, although able to grow at a temperature of 20°–30° C, require a short period at 60° C (140° F) for their spores to germinate.

Symbionts

Associations. Among symbiotic fungi, those that enter into mycorrhizal relationships and the lichens (see below *Lichens*) are probably the best-known. Another kind of symbiotic relationship between fungi and host organisms is found between fungi and certain insects.

A large number of fungi "infect" the roots of plants, forming an association with them called mycorrhiza (plural mycorrhizae). This association differs markedly from ordinary root infection, which is responsible for root rot diseases. It is a non-disease-producing association in which the fungus invades the root and derives nutrients from it. Mycorrhizal fungi establish a mild form of parasitism that in many instances verges on mutualism; *i.e.*, it is beneficial to the plant as well as to the fungus.

There are two types of mycorrhizae, ectomycorrhizae and endomycorrhizae. Ectomycorrhizae are fungi that are only externally associated with the plant root, while the endomycorrhizae form their associations within the cells of the host.

Among the mycorrhizal fungi are the boletes (Basidiomycetes, family Boletaceae), whose mycorrhizal relationships with larch trees (*Larix*) and other conifers have long been known. Others include the truffles (Ascomycetes, order Tuberales), some of which are believed to form mycorrhizae with oak (*Quercus*) or beech (*Fagus*) trees, and various species of *Rhizoctonia* (mycelial stage of certain Basidiomycetes), which form orchid mycorrhizae.

Insect associations

The symbiotic relationship between certain fungi and insects is exemplified by that of the fungal genus *Septobasidium* and scale insects (order Homoptera) that feed on trees. The mycelium forms elaborate structures over colonies of insects feeding on the bark. Each insect sinks its proboscis (tubular sucking organ) into the bark and remains there the rest of its life, sucking sap. The fungus sinks haustoria (special absorbing structures) into the bodies of some of the insects and feeds on them without killing them. The parasitized insects are, however, rendered sterile.

The perpetuation of the insect species and the spread of the fungus are accomplished by the uninfected members of the colony, which live in fungal "houses," safe from enemies. Newly hatched scale insects crawl over the surface of the fungus, which is at that time sporulating. Fungal spores adhere to the young insects, germinate, and infect them. As the young insects settle down in a new place on the bark to begin feeding, they establish new fungal colonies. Thus, part of the insect colony is sacrificed to the fungus as food in return for the fungal protection provided for the rest of the insects. The insect is parasitic on the tree and the fungus is parasitic on the insect, but the tree is the ultimate victim.

The sooty molds constitute another interesting ecological group of fungi associated with insects. The majority of these are tropical or subtropical, but some species occur in the temperate zones. All sooty molds are epiphytic (*i.e.*, they grow on the surfaces of other plants), but only in areas where scale insects are present. The fungi parasitize neither the plants nor the insects but rather obtain their nourishment exclusively from the honeydew secretions of the scale insects. Growth of the dark mycelium over the plant leaves, however, is often so dense as to reduce greatly the intensity of the light that reaches the leaf surface; this reduction in turn significantly reduces the rate of photosynthesis.

FORM AND FUNCTION

Septate and nonseptate hyphae

Structure of the thallus. In almost all fungi the hypha, and therefore the thallus—the undifferentiated, nutrient-absorbing body of the fungus—has cell walls. (The thalli of the true slime molds lack cell walls and, for this and other reasons, are classified as protists rather than fungi.) A hypha is a multibranched tubular cell filled with cytoplasm. The tube itself may be either continuous throughout or divided into compartments, or cells, by cross walls called septa (singular septum). In nonseptate (*i.e.*, coenocytic) hyphae the nuclei are scattered throughout the cytoplasm. In septate hyphae each cell may contain one to many nuclei, depending on the type of fungus or the stage of hyphal development. The cells of fungi are similar in structure to those of many other organisms. The minute nucleus, readily seen only in young portions of the hypha, is surrounded by a double membrane and typically contains one nucleolus. In addition to the nucleus, various organelles, such as the endoplasmic reticulum, the Golgi apparatus, ribosomes, and liposomes, are scattered throughout the cytoplasm.

Hyphae usually are either nonseptate (generally in the more primitive fungi) or incompletely septate. This permits the movement of cytoplasm (cytoplasmic streaming) from one cell to the next. Cytoplasmic streaming, however, is movement in only one direction (toward the growing end of the hypha), and it is movement that does not effect locomotion. In those septae that are nonperforated, thin cytoplasmic strands called plasmodesmata can develop outside the hyphal wall between adjacent cells.

Variations in the structure of septae are numerous in the fungi. Some fungi have sievelike septae, called pseudosepta, while fungi in other groups have septae with one to few pores which are small enough in size that they regulate the movement of nuclei and cytoplasm to adjacent cells. Many of the Basidiomycetes have a somewhat characteristic septal structure, called the dolipore septum, composed of a septal pore cap surrounding a septal swelling and septal pore. This organization permits cytoplasm and small organelles to pass through but restricts the movement of nuclei to varying degrees, forcing the nucleus to constrict as it passes through.

The hyphal wall

The wall of the hypha is complex in both composition and structure. Its exact chemical composition varies in different fungal groups. In some fungi the wall contains considerable quantities of cellulose, a complex carbohydrate that is the chief constituent of the cell walls of plants; in most fungi, however, two different polymers, chitin and a special β-1,3-1,6-glucan (a polymer of glucose linked at the third carbon and branched at the sixth), are the main structural components of the wall. In a few fungi, both chitin and cellulose are components of the cell wall. Among the many other chemical substances in the walls of fungi are some that may thicken or toughen the wall of tissues, thus imparting rigidity and strength. The chemical composition of the wall of a particular fungus may vary at different stages of the organism's growth—a possible indication that the wall plays some part in determining the form of the fungus. In some yeasts, fusion of sexually functioning cells is brought about by the interaction of specific chemical substances on the walls of two compatible mating types.

When the mycelium grows in or on a surface—the soil, a log, a culture medium—it appears as a mass of loose, cottony threads. The richer the composition of the growth medium, the more profuse the threads and the more feltlike the mass. On the sugar-rich growth substances used in laboratories, the assimilative (somatic) hyphae are so interwoven as to form a thick, almost leathery colony. On the soil, inside a leaf, in the skin of animals, or in other parasitized plant or animal tissues, the hyphae are usually spread in a loose network. The mycelium of the so-called higher fungi does, however, become organized at times into compact masses of different sizes that serve various functions. Some of these masses, called sclerotia, become extremely hard and serve to carry the fungus over periods of adverse conditions of temperature and moisture. Ergot is an example. The underground sclerotia of *Poria cocos*, an edible pore fungus known in the United States under the Indian name tuckahoe, may reach a diameter of 20 to 25 centimetres. Various other tissues are also produced by the interweaving of the assimilative hyphae of some fungi. Stromata (singular stroma) are cushionlike tissues that bear spores in various ways. Rhizomorphs are long strands of parallel hyphae cemented together. Those of the honey mushroom *Armillariella mellea*, which are black and resemble shoestrings, are intricately constructed and are differentiated to conduct water and food materials from one part of the thallus to another.

Sporophores and spores. When the mycelium of a fungus reaches a certain stage of growth, it begins to produce spores either directly on the somatic hyphae or, more often, on special sporiferous (spore-producing) hyphae, which may be loosely arranged or grouped into intricate structures called the fruiting bodies or sporophores. The type of sporophore produced is characteristic of the group to which a fungus belongs, and the classification of fungi is based almost entirely on the characters of their sporophores and spores.

The lower (*i.e.*, more primitive) fungi produce spores in sporangia, which are saclike sporophores whose entire cytoplasmic contents cleave into spores, called sporangiospores. Thus, they differ from more advanced fungi in that their asexual spores are endogenous. Sporangiospores are either naked and flagellated (zoospores) or walled and nonmotile (aplanospores). The more primitive aquatic and terrestrial fungi tend to produce zoospores. The zoospores of aquatic fungi swim in the surrounding water by means of one or two variously located flagella (whiplike organs of locomotion), depending on the group to which the fungus belongs. Zoospores produced by terrestrial fungi are released after a rain from the sporangia in which they are borne and swim for a time in the rainwater between soil particles or on the wet surfaces of plants, where the sporangia are formed by parasitic fungi. After some time, the zoospores lose their flagella, surround themselves with walls, and encyst. Each cyst germinates by producing a germ tube. The germ tube may develop a mycelium or a reproductive structure, depending on species and environmental conditions. The bread molds (Zygomycetes), which are the most advanced of the lower fungi, produce only aplanospores (nonmotile spores) in their sporangia.

More advanced fungi

The more advanced fungi (Ascomycetes, Deuteromycetes, and Basidiomycetes) do not produce motile spores of any kind even though some of them are aquatic in fresh or marine waters. In these fungi, asexually produced spores (usually called conidia) are produced exogenously, typically formed terminally or laterally on special spore-producing hyphae, the conidiophores, which are variously arranged—*i.e.*, singly on the hyphae or grouped in special asexual fruiting bodies such as flask-shaped pycnidia, mattresslike acervuli, cushion-shaped sporodochia, or sheaflike synnemata.

Sexually produced spores of the higher fungi result from meiosis and are formed either in saclike structures (asci) typical of the Ascomycetes or on the surface of typically club-shaped structures (basidia) typical of the Basidiomycetes. Asci and basidia may be borne naked, directly on the hyphae, or in various types of sporophores, called ascocarps or basidiocarps, depending on whether they bear asci or basidia, respectively. Well-known examples of ascocarps are the morels, the cup fungi, and the truffles. Commonly encountered basidiocarps are mushrooms, puffballs, stinkhorns, and bird's-nest fungi.

Growth. Under favourable environmental conditions, fungal spores germinate and form hyphae. During this process, the spore absorbs water through its wall, the cytoplasm becomes activated, nuclear division takes place, more cytoplasm is synthesized, and from the wall of the spore a germ tube bulges out, enveloped by a wall of its own that is formed as the tube grows.

The hypha may be roughly divided into three regions: (1) the apical zone about 5–10 micrometres (0.0002–0.0004 inch) in length, (2) the subapical region, extending about 40 micrometres back of the apical zone, which is rich in cytoplasmic components, such as nuclei, Golgi apparatus, ribosomes, mitochondria, the endoplamsic reticulum, and vesicles, but is devoid of vacuoles, and (3) the zone of vac-

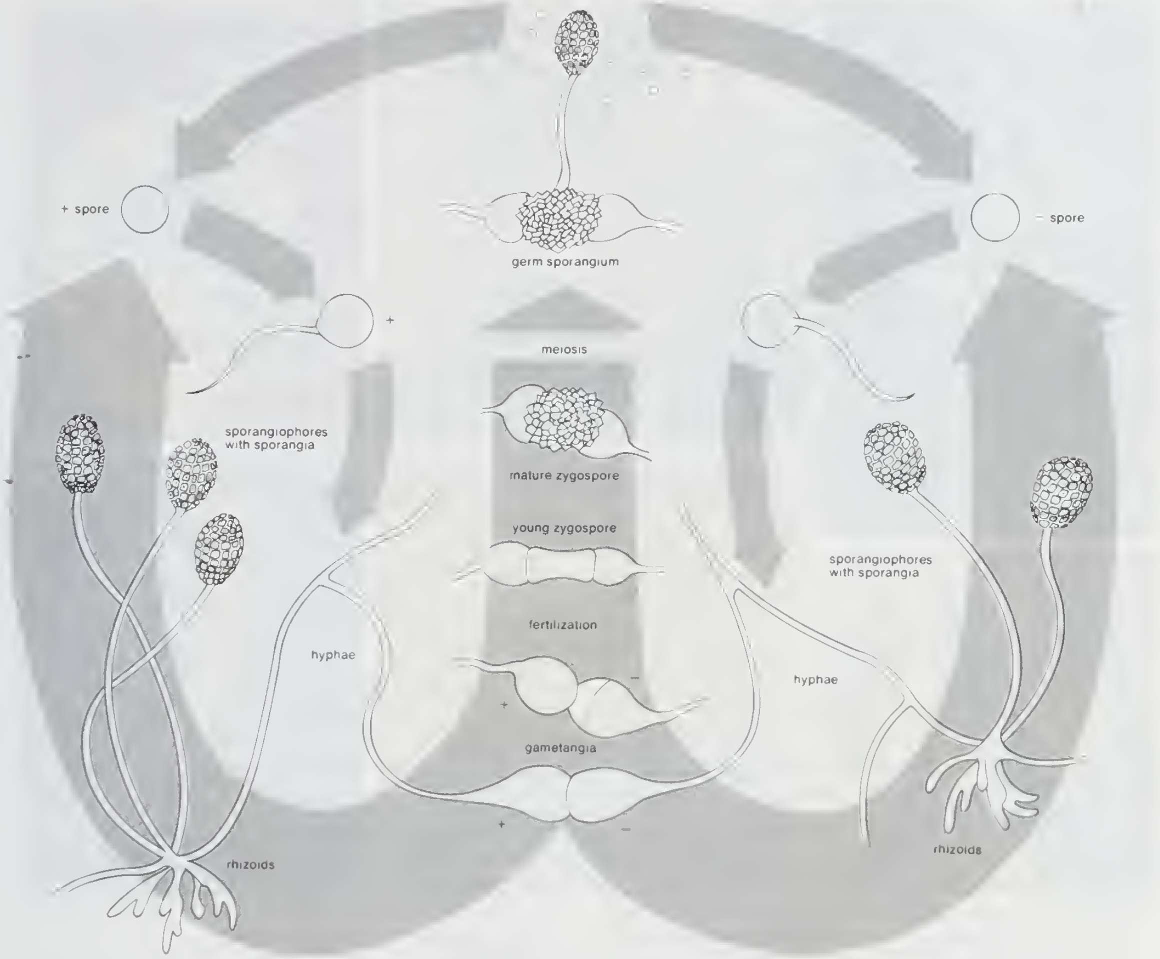


Figure 5: Life cycle of the common bread mold *Rhizopus stolonifer*, a zygomycete.

From C J Alexopoulos, *Introductory Mycology* 2nd ed copyright © (1962) John Wiley and Sons Inc reprinted by permission

uolation, which is characterized by the presence of many vacuoles and the accumulation of lipids.

Growth of hyphae takes place almost exclusively in the apical zone (*i.e.,* at the very tip). This is the region where the cell wall extends continuously to produce a long hyphal tube. The cytoplasm within the apical zone is filled with numerous vesicles. These bubblelike structures are usually too small to be seen with an ordinary microscope but are clearly evident under the electron microscope. In higher fungi (Ascomycetes and Basidiomycetes), the apical vesicles can be detected with an ordinary microscope equipped with special optics (phase contrast), as a round spot with a somewhat diffuse boundary. This body is universally known by its German name, the Spitzenkörper, and its position determines the direction of growth of a hypha.

The growing tip eventually gives rise to a branch. This is the beginning of the branched mycelium. Growing tips that come in contact with neighbouring hyphae often fuse with them to form a hyphal net. In such a vigorously growing system, the cytoplasm is in constant motion, streaming toward the growing tips. Eventually, the older hyphae become highly vacuolated and may be stripped of most of their cytoplasm. All living portions of a thallus are potentially capable of growth. If a small piece of mycelium is placed under conditions favourable for growth, it develops into a new thallus even if no growing tips are included in the severed portion.

Growth of a septate mycelium (*i.e.,* with cross walls between adjacent cells) entails the formation of new septa in the young hyphae. Septa are formed by ringlike growth from the wall of the hypha toward the centre until the septa are complete. In the higher fungi the septum stops growing before it is complete; the result is a central pore through which the cytoplasm flows, thus establishing organic connection throughout the thallus.

The individual fungus is potentially immortal, for it continues to grow at the hyphal tips as long as conditions remain favourable. It is possible that, in undisturbed places, mycelia exist that have grown continuously for many thousands of years. The older parts of the hyphae die and decompose, releasing nitrogen and other nutrients into the soil.

Nutrition. Unlike green plants, which use carbon dioxide and light as sources of carbon and energy, respectively, fungi meet these two requirements by assimilating preformed organic matter; carbohydrates are the preferred nutrient source. Fungi can readily absorb and metabolize a variety of soluble carbohydrates, such as glucose, xylose, sucrose, and fructose, but are also characteristically well equipped to use insoluble carbohydrates like starches, cellulose, hemicelluloses, and lignin. To do so, they must first digest these polymers extracellularly. Saprobic fungi obtain their food from dead organic material; parasitic fungi do so by feeding on living organisms (usually plants), thus causing disease.

External digestion of food

Fungi secure food through the action of enzymes (biological catalysts) secreted into the surface on which they are growing; the enzymes digest the food, which then is absorbed directly through the hyphal walls. The rotting of fruits, such as peaches and citrus fruits in storage, demonstrates this phenomenon, in which the infected parts are softened by the action of the fungal enzymes. In brown rot of peaches, the softened area is somewhat larger than the actual area invaded by the hyphae: the periphery of the brown spot has been softened by enzymes that act ahead of the invading mycelium. Food must enter the hyphae in solution, and, since most fungi have no special absorbing organs, the entire mycelial surface is capable of taking in materials dissolved in water. Some fungi, however, produce special rootlike hyphae, called rhizoids, which anchor the thallus to the growth surface and proba-

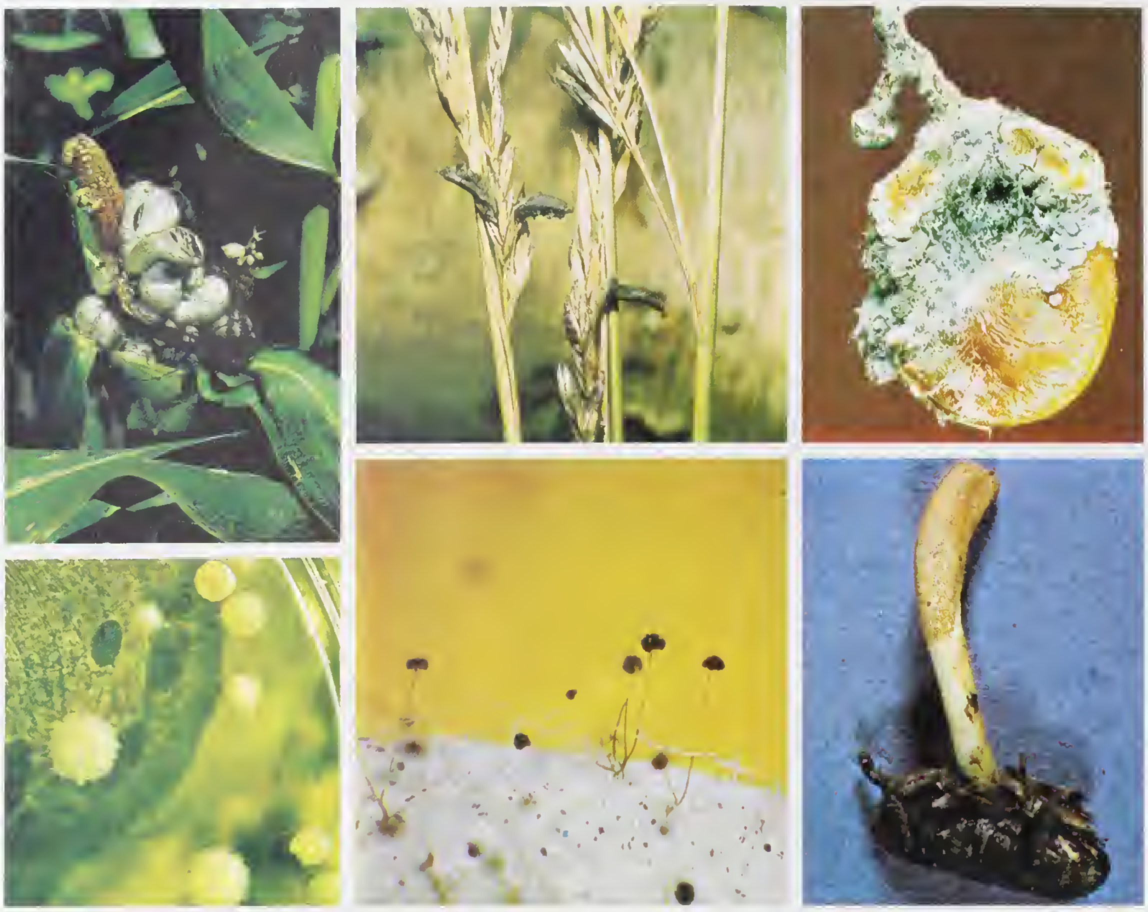


Figure 6: *Fungi of economic or scientific interest.*
(Top left) Corn smut (*Ustilago*) on *Zea mays* yearly destroys a considerable percentage of the corn (maize) crop. (Top centre) Ergot (*Claviceps purpurea*), the original source of LSD-25 as well as a drug used in obstetrics, attacks rye and other cereal grains and causes the disease ergotism when eaten. (Top right) Blue-green mold (*Penicillium*), shown on grape, attacks many fruits and is the source of the drug penicillin. (Bottom left) Brewer's yeast colonies (*Saccharomyces cerevisiae*) growing on agar medium. (Bottom centre) Bread mold (*Rhizopus stolonifer*) showing the sporangia that bear sporangiospores (asexual spores). (Bottom right) Parasitic *Cordyceps millitaris* attacking insect pupa.

(Top left) J C Allen and Son, (top centre, top right) Walter Dawn, (bottom left) A to Z Botanical Collection—EB Inc , (bottom centre) Grant Heilman—EB Inc , (bottom right) F K Anderson—EB Inc

bly also absorb food. Many parasitic fungi are even more specialized in this respect, producing special absorptive organs called haustoria.

Saprobiosis. Together with the bacteria, saprobic fungi are to a large extent responsible for the decomposition of organic matter. They are also responsible for the decay and decomposition of foodstuffs. Among other destructive saprobes are fungi that destroy timber and timber products as their mycelia invade and digest the wood; many of these produce their spores in large, woody, fruiting bodies—*e.g.,* bracket or shelf fungi. Paper, textiles, and leather are often attacked and destroyed by fungi. This is particularly true in tropical regions, where temperature and humidity are often very high.

The nutritional requirements of saprobes (and of some parasites that can be cultivated artificially) have been determined by growing fungi experimentally on various synthetic substances of known chemical composition. Fungi usually exhibit the same morphological characteristics in these culture media as they do in nature. Carbon is supplied in the form of sugars or starch; the majority of fungi thrive on such sugars as glucose, fructose, mannose, maltose, and, to a lesser extent, sucrose. Decomposition products of proteins—*e.g.,* proteoses, peptones, and amino acids—can be used by most fungi as nitrogen sources; ammonium compounds and nitrates also serve for many species. It is doubtful, however, that any fungus can combine, or fix, atmospheric nitrogen into usable compounds. Chemical elements such as phosphorus, sulfur, potassium, magnesium, and small quantities of iron, zinc, manganese, and copper are needed by most fungi for vigorous growth; elements such as calcium, molybdenum, and gallium are required by at least some species. Oxygen and hydrogen are absolute requirements; they are supplied in the form of water or are obtained from carbohydrates. Many fungi, deficient in thiamine and biotin, must obtain these vitamins from the environment; most fungi appear able to synthesize all other vitamins necessary for their growth and reproduction. As a rule, fungi are aerobic organisms; *i.e.,* they require free oxygen. Fermentations, however, take place under anaerobic conditions as well. Knowledge of the physiology of saprobic fungi has enabled industry to use several species for fermentation purposes.

Parasitism. In contrast with the saprobic fungi, parasitic fungi attack living organisms, penetrate their outer defenses, invade them, and obtain nourishment from living cytoplasm, causing disease and sometimes the death of the host. Most pathogenic (disease-causing) fungi are parasites of plants, but several are known to cause diseases of humans and lower animals. Most parasites enter the host through a natural opening, such as a stomate (microscopic air pore) in a leaf, a lenticel (small opening through bark) in a stem, a broken plant hair or a hair socket in a fruit, or a wound in the plant or animal epidermis (skin). Such wounds may be insect punctures or accidentally inflicted scratches, cuts, or bruises. Among the most common and widespread diseases of plants caused by fungi are the various downy mildews (*e.g.,* of grape, onion, tobacco), the powdery mildews (*e.g.,* of grape, cherry, apple, peach, rose, lilac), the smuts (*e.g.,* of corn, wheat, onion), the rusts (*e.g.,* of wheat, oats, beans, asparagus, snapdragon, hollyhock), apple scab, brown rot of stone fruits, and

Fungus-caused diseases

various leaf spots, blights, and wilts. These diseases cause great damage annually throughout the world.

Infection of a plant takes place when the spores of a pathogenic fungus fall on the leaves or the stem of a susceptible host and germinate, each producing a germ tube. The tube grows on the surface of the host until it finds an opening; then the tube enters the host, puts out branches between the cells of the host, and forms a mycelial network within the invaded tissue. The germ tubes of some fungi produce special pressing organs called appressoria, from which a microscopic, needlelike peg presses against and punctures the epidermis of the host; after penetration, a mycelium develops in the usual manner. Many parasitic fungi absorb food from the host cells through the hyphal walls appressed against the cell walls of the host's internal tissues. Others produce haustoria (special absorbing structures) that branch off from the intercellular hyphae and penetrate the cells themselves. Haustoria, which may be short, bulbous protrusions or large branched systems filling the whole cell, are characteristically produced by obligate (*i.e.,* invariably parasitic) parasites; some facultative (*i.e.,* occasionally parasitic) parasites such as the potato blight, *Phytophthora infestans,* also produce them. Obligate parasites, which require living cytoplasm and have extremely specialized nutritional requirements, are exceptionally difficult, and often impossible, to grow. Examples of obligate parasites are the downy mildews, the powdery mildews, and the rusts.

Predation. A number of fungi have developed ingenious mechanisms for trapping microorganisms, such as amoebas, roundworms (nematodes), and rotifers. After the prey is captured, the fungus penetrates its body and quickly destroys it. Many of these fungi secrete adhesive substances over the surface of their hyphae, so that a passing animal that touches any portion of the mycelium adheres firmly to it. Soon a penetration tube grows out of a hypha and penetrates the host's soft body. This haustorium grows and branches, and enzymes secreted by it quickly kill the animal, whose cytoplasm serves as food for the fungus.

Animal-trapping fungi

Other fungi produce hyphal loops in which small animals become ensnared until the fungus is able to send haustoria into their bodies and kill them. Perhaps the most amazing of these fungal traps are the so-called constricting rings of some species of *Arthrobotrys, Dactylella,* and *Dactylaria*—soil-inhabiting fungi easily grown under laboratory conditions. In the presence of nematodes, the mycelium produces large numbers of rings through which the average nematode is barely able to pass. When a nematode rubs the inner wall of a ring, which usually consists of three cells with touch-sensitive inner surfaces, the cells of the ring swell rapidly, and the resulting constriction holds the worm tightly. All efforts of the nematode to free itself fail, and a hypha, which grows out of one of the swollen ring cells at its point of contact with the worm, penetrates the animal's body, branches therein, and kills the host, which is then used for food by the fungus. In the absence of nematodes, these fungi do not usually produce rings in appreciable quantities. A substance (nemin) of largely unknown chemical composition is secreted by the nematodes and stimulates the fungus to form the mycelial rings. (C.J.Al./Ed.)

Lichens. *General features.* A lichen is an association between a fungus and an alga that results in a form distinct from either symbiont. Although lichens appear to be single plants, under a microscope the association is seen to consist of millions of cells of algae (called the phycobiont) woven into a matrix formed of the filaments of the fungus (called the mycobiont). The majority of mycobionts are placed in a single group of Ascomycetes called the Lecanorales, which are characterized by an open, often button-shaped fruit called an apothecium. The remaining mycobionts are distributed among various fungal groups—*e.g.,* Sphaeriales, Caliciales, Myrangiales, Pleosporales, and Hysteriales. Although there are various types of phycobionts, most of them also belong to a single group; *i.e.,* half the lichen associations contain species of *Trebouxia,* a single-celled green alga.

Authorities have not been able to establish with any certainty when and how these associations evolved, although lichens must have evolved more recently than their components and probably arose independently from different groups of fungi and algae. It seems, moreover, that the ability to form lichens can spread to new groups of fungi and algae. Lichens are a biological group lacking formal status in the taxonomic framework of living organisms. Although the mycobiont and phycobiont have Latin names, the product of their interaction, a lichen, does not. Earlier names given to lichens as a whole now are considered names for the fungus alone. Classification of lichens is difficult and remains controversial. Part of the problem is that the taxonomy of lichens was established before their dual nature was recognized; *i.e.,* the association was treated as a single entity.

Taxonomic status

Approximately 15,000 different kinds of lichens, some of which provide forage for reindeer and products for humans, have been described. Some lichens are leafy and form beautiful rosettes on rocks and tree trunks; others are filamentous and drape the branches of trees, sometimes reaching a length of 2.75 metres (9 feet). At the opposite extreme are those smaller than a pin head and seen only with a magnifying lens. Lichens grow on almost any type of surface and can be found in almost all areas of the world. They are especially prominent in bleak, harsh regions where few plants can survive. They grow farther north and farther south and higher on mountains than most plants.

The thallus of a lichen has one of several characteristic growth forms (crustose, foliose, or fruticose—see below *Form and function*). Crustose thalli, which resemble a crust closely attached to a surface, are drought-resistant and well adapted to dry climates. They prevail in deserts, Arctic and Alpine regions, and ice-free parts of Antarctica. Foliose, or leafy, thalli grow best in areas of frequent rainfall; two foliose lichens, *Hydrothyria venosa* and *Dermatocarpon fluviatile,* grow on rocks in freshwater streams of North America. Fruticose (stalked) thalli and filamentous forms prefer to utilize water in vapour form and are prevalent in humid, foggy areas such as seacoasts and mountainous regions of the tropics.

Humans have used lichens as food, as medicine, and in dyes. A versatile lichen of economic importance, *Cetraria islandica,* commonly called Iceland moss, sometimes is used either as an appetite stimulant or as a foodstuff in reducing diets; it has also been mixed with bread. Iceland moss also has been used to treat diabetes, nephritis, and catarrh. Lichens have little medical value. One lichen, *Lecanora esculenta,* is reputed to be the manna that in ancient days fell from the skies and served as a food source for humans and domestic animals.

Lichens as dye sources

Lichens are well known as dye sources. Dyes derived from them have an affinity for wool and silk and are formed by decomposition of certain lichen acids and conversion of the products. One of the best-known lichen dyes is orchil, which has a purple or red-violet colour. Orchil-producing lichens include species of *Ochrolechia, Roccella,* and *Umbilicaria.* Litmus, formed from orchil, is widely used as an acid-base indicator. Synthetic coal tar dyes, however, have replaced lichen dyes in the textile industry, and usage of orchil now is limited to its use as a food-colouring agent and an acid-base indicator. A few lichens (*e.g., Evernia prunastri*) are used in the manufacture of perfumes.

Caribou and reindeer depend on lichens for two-thirds of their food supply. In northern Canada an acre of land undisturbed by animals for 120 years or more may contain 250 kilograms (560 pounds) of lichens; some forage lichens that form extensive mats on the ground are *Cladonia alpestris, C. mitis, C. rangiferina,* and *C. sylvatica.* Arboreal lichens such as *Alectoria, Evernia,* and *Usnea* also are valuable as forage. An acre of mature black spruce trees in northern Canada, for example, may contain more than 270 kilograms of lichens on branches within 3 metres of the ground.

Form and function. Although the fungal symbionts of many lichens have fruiting structures on or within their thalli and may release numerous spores that develop into fungi, indirect evidence suggests that natural unions of fungi and algae occur only rarely among some lichen groups, if indeed they occur at all. In addition, free-living, potential phycobionts are not widely distributed; *e.g.*, despite repeated searches, free-living populations of *Trebouxia* have not been found. This paradox, an abundance of fungal spores and a lack of algae capable of forming associations, implies that the countless spores produced by lichen fungi are functionless, at least so far as propagation of the association is concerned. Some phycobionts—*i.e.*, species of *Nostoc* and *Trentpohlia*—can exist as free-living populations, so that natural reassociations could occur in a few lichens.

Some lichens have solved or bypassed the problem of recombination. In a few lichens (*e.g., Endocarpon, Staurothele*) algae grow among the tissues of a fruiting body and are discharged along with fungal spores; such phycobionts are called hymenial algae. When the spores germinate, the algal cells multiply and gradually form lichens with the fungus. Other lichens form structures, especially soredia, that are effective in distributing the association. A soredium, consisting of one or several algal cells enveloped by threadlike fungal filaments, or hyphae, may develop into a thallus under suitable conditions. Lichens without soredia may propagate by fragmentation of their thalli. Many lichens develop small thalloid extensions, called isidia, that also may serve in asexual propagation if broken off from the thallus.

In addition to these mechanisms for propagation, the individual symbionts have various methods of reproduction. Ascolichens—*i.e.*, lichens in which the mycobiont is an ascomycete—for example, form fruits (ascocarps) similar to those of free-living ascomycetes, except that the mycobiont's fruits are capable of producing spores for a longer period of time. The algal symbiont within the lichen thallus reproduces by the same methods as its free-living counterpart.

Most lichen phycobionts are penetrated to varying degrees by specialized fungal structures called haustoria. *Trebouxia* lichens have a pattern in which deeply penetrating haustoria are prevalent in associations lacking a high degree of thalloid organization. On the other hand, superficial haustoria prevail among forms with highly developed thalli. *Lecanora* and *Lecidea*, for example, have individual algal cells with as many as five haustoria that may extend to the cell centre. *Alectoria* and *Cladonia* have haustoria that do not penetrate far beyond the algal cell wall. A few phycobionts, such as *Coccomyxa* and *Stichococcus*, which are not penetrated by haustoria, have thin-walled cells that are pressed close to fungal hyphae.

Benefits of symbiosis

The flow of nutrients and metabolites between the symbionts is the basic foundation of the symbiotic system. A simple carbohydrate formed in the algal layer eventually is excreted, taken up by the mycobiont, and transformed into a different carbohydrate. The release of carbohydrate by the phycobiont and its conversion by the mycobiont occur rapidly. Whether or not the fungus influences the release of carbohydrate by the alga is not known with certainty, but it is known that carbohydrate excretion by the alga decreases rapidly if it is separated from the fungus.

Carbohydrate transfer is only one aspect of the symbiotic interaction in lichens. The alga may provide the fungus with vitamins, especially biotin and thiamine, important because most lichen fungi that are grown in the absence of algae have vitamin deficiencies. The alga also may contribute a substance that causes structural changes in the fungus since it forms the typical lichen thallus only in association with an alga.

One contribution of the fungus to the symbiosis concerns absorption of water vapour from the air; the process is so effective that, at high levels of air humidity, the phycobionts of some lichens photosynthesize at near-maximum rates. The upper region of a thallus provides shade for the underlying algae, some of which are sensitive to strong light. In addition, the upper region may contain pigments or crystals that further reduce light intensity and act as filters, absorbing certain wavelengths of light.

Lichens synthesize a variety of unique organic compounds that tend to accumulate within the thallus; many of these substances are coloured and are responsible for the red, yellow, or orange colour of lichens.

Thalloid types

A lichen thallus or composite body has one of two basic structures. In a homoiomerous thallus, the algal cells, distributed throughout the structure, are more numerous than those of the fungus. The more common type of thallus, a heteromerous thallus, has four distinct layers, three of which are formed by the fungus and one by the alga. The fungal layers are called upper cortex, medulla, and lower cortex. The upper cortex consists of either a few layers of tightly packed cells or hyphae that may contain pigments. A cuticle may cover the cortex. The lower cortex, which is similar in structure to the upper cortex, participates in the formation of attachment structures called rhizines. The medulla, located below the algal layer, is the widest layer of a heteromerous thallus. It has a cottony appearance and consists of interlaced hyphae. The loosely structured nature of the medulla provides it with numerous air spaces and allows it to hold large amounts of water. The algal layer, about three times as wide as a cortex, consists of tightly packed algal cells enveloped by fungal hyphae from the medulla.

A heteromerous thallus may have a stalked (fruticose), crustlike (crustose), or a leafy (foliose) form; many transitional types exist. It is not known, moreover, which growth form is primitive and which is advanced. Fruticose lichens, which usually arise from a primary thallus of a different growth form (*i.e.*, crustose, foliose), may be shrubby or pendulous or consist of upright stalks. The fruticose form usually consists of two thalloid types: the primary thallus is crustlike or lobed; the secondary thalli, which originate from the crust or lobes of the primary thallus, consists of stalks that may be simple, cup-shaped, intricately branched, and capped with brown or red fruiting bodies called apothecia. Fruticose forms such as *Usnea* may have elongated stalks with a central solid core that provides strength and elasticity to the thallus.

The crustose thallus is in such intimate contact with the surface to which it is attached that it usually cannot be removed intact. Some crustose lichens grow beneath the surface of bark or rock so that only their fruiting structures penetrate the surface. Crustose lichens may have a hypothallus—*i.e.*, an algal-free mat of hyphae extending beyond the margin of the regular thallus. Crustose form varies: granular types such as *Lepraria*, for example, have no organized thalloid structure; but some *Lecanora* species have highly organized thalli, with lobes that resemble foliose lichens lacking a lower cortex.

From *An Evolutionary Survey of the Plant Kingdom* by Robert F. Scagel *et al.*, © 1965 by Wadsworth Publishing Company, Inc., Belmont, California 94002, reprinted by permission of the publisher

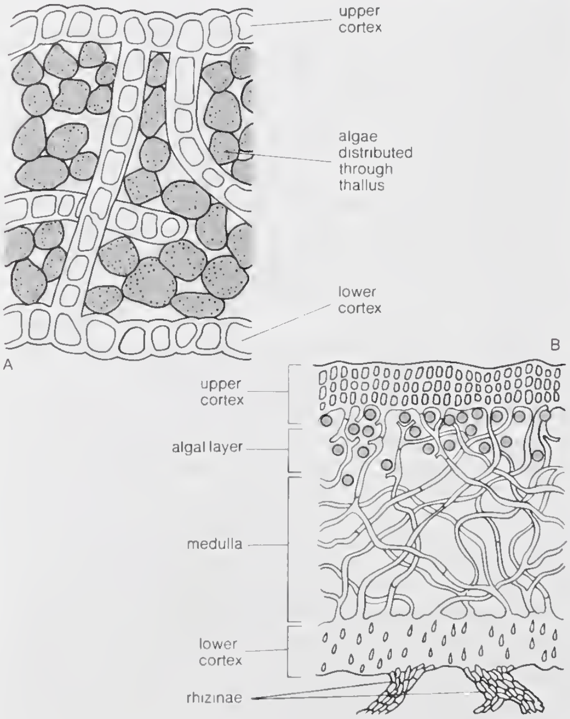


Figure 7: *Distinctions based on distribution of algal cells.* (A) Homoiomerous thallus; (B) heteromerous thallus.

The foliose forms are flat, leaflike, and loosely attached to a surface. The largest known lichens have a foliose form; species of *Sticta* may attain a diameter of about a metre. Other common foliose genera include *Cetraria, Parmelia, Peltigera,* and *Physcia. Umbilicaria,* called the common rock tripe, differs from other foliose forms in its mode of attachment in that its platelike thallus attaches at the centre to a rock surface.

Fruiting bodies and spores

The complex fruiting bodies (ascocarps) of lichen fungi are of several types. The factors that induce fruiting in lichens have not been established with certainty. Spores of lichen fungi (ascospores) are of extremely varying sizes and shapes; *e.g., Pertusaria* has one or two large spores in one ascus (saclike bodies containing the ascospores), and *Acarospora* may have several hundred small spores per ascus. Although in most species the ascospore generally has one nucleus, it may be single-celled or multicellular, brown or colourless; the *Pertusaria* spore, however, is a single cell containing 200 nuclei. Another type of fungal spore may be what are sometimes called spermatia (male fungal sex cells) or pycnidiospores; it is not certain that these structures have the ability to germinate and develop into a fungal colony. Few lichen fungi produce conidia, a type of asexual spore common among ascomycetes.

The metabolic activity of lichens is greatly influenced by the water content of the thallus. The rate of photosynthesis may be greatest when the amount of water in the thallus is from 65 to 90 percent of the maximum. During drying conditions, the photosynthetic rate decreases and below 30 percent is no longer measurable. Although respiration also decreases rapidly below 80 percent water content, it persists at low rates even when the thallus is air-dried. Since lichens have no mechanisms for water retention or uptake from the surface to which they are attached, they very quickly lose the water vapour they absorb from the air. The rapid drying of lichens is a protective device; *i.e.,* a moisture-free lichen is more resistant to temperature and light extremes than is a wet one. Frequent drying and wetting of a thallus is one of the reasons lichens have a slow growth rate.

Maximum photosynthesis in lichens takes place at temperatures of 15° to 20° C. More light is needed in the spring and summer than in the winter. The photosynthetic apparatus of lichens is remarkably resistant to cold temperatures. Even at temperatures below 0° C, many lichens can absorb and fix considerable amounts of carbon dioxide. Respiration is much less at low temperatures so that, in nature, the winter months may be the most productive ones for lichens. (V.A./Ed.)

EVOLUTION AND PHYLOGENY

The origin of the fungi is obscure, the fossil record being scanty and virtually meaningless. The older theory supposed the fungi to have originated by loss of chlorophyll from one or two groups of algae, one school favouring the development of the fungi monophyletically (*i.e.,* from a single ancestor) from the green algae, the other postulating that the lower fungi originated from the green algae but that the Ascomycetes came from the class Florideae of the red algae and later gave rise to the Basidiomycetes. Most present-day mycologists derive the fungi from ancestral flagellates (algal or protozoan organisms bearing flagella, whiplike swimming organs) but yield that the Oomycetes may belong to a different evolutionary line because of their unique biochemical and cytological features: they synthesize the amino acid lysine, they have cellulosic walls and a special organization of the tryptophan-synthesizing enzymes; their thallus is diploid; and they reproduce through oogamy (production of differentiated egg cells).

Taxonomic criteria

As for the interrelationships among fungal groups, there is much controversy. The modern tendency is to emphasize flagellation as an important phylogenetic criterion in the lower fungi. Thus, all the posteriorly uniflagellate fungi (*i.e.,* those with a single flagellum located at the rear end of the organism) are brought together in one class, Chytridiomycetes, all the anteriorly uniflagellate in the class Hyphochytridiomycetes, and so on, as may be seen in the section on classification below. Whether the Plasmodiophoromycetes should be grouped with the Myxomycetes (slime molds) and removed from the fungi because of their plasmodial phase and their swarm cells, which bear two anterior whiplash flagella, is a moot question.

Biochemical characters have been useful markers to map the probable evolutionary relationships of fungi. Because of common biochemical attributes, such as similarity in wall composition (presence of both chitin and β-1,3-1,6-glucan), the same pattern of organization of tryptophan enzymes, and synthesis of lysine by the same unique pathway (aminoadipic acid), the Chytridiomycetes, Ascomycetes, and Basidiomycetes are believed to constitute the main axis of fungal evolution, with the flagellated Chytridiomycetes representing the most primitive or ancestral forms. Other major groups of fungi, Zygomycetes (mucorales), Hemiascomycetes (ascomycetous yeasts), and Heterobasidiomycetes (basidiomycetous yeasts), are also in the same evolutionary camp; they have chitinous walls and make lysine by the aminoadipic acid pathway but show significant differences in cell-wall composition and organization of tryptophan biosynthetic enzymes. Hence, these fungal groups are believed to be side branches from the main evolutionary axis.

Virtually all mycologists agree that the Basidiomycetes have been derived from the Ascomycetes. This opinion is based on the similarity of the nuclear cycle of the ascus and basidium, the supposed homology of the clamp connection (a structure joining two adjacent cells in the Basidiomycetes) with the crozier (a hook-shaped terminal cell found in Ascomycetes), and the similarity of the binucleate mycelium of Basidomycetes with the ascogenous (ascus-producing) hyphae of Ascomycetes. Not all, however, are agreed as to which group of Ascomycetes gave rise to the Basidiomycetes, nor indeed as to which Basidiomycetes are primitive and which are advanced. Whether the holobasidium (simple, club-shaped basidium) or the heterobasidium (septate or deeply divided basidium) came first is a highly controversial question that has a great bearing on the origin of the Basidiomycetes. Also, the fact that the dolipore (inflated) septum has not been found in the rusts or smuts or in any clearly nonbasidiomycetous group poses an interesting question on the origin of the Basidiomycetes.

CLASSIFICATION

Distinguishing taxonomic features. The fungi as a group are distinguished from other organisms by the nature of their somatic (body) and reproductive structures and by the mode of nutrition they employ. Within the division Mycota, the classes are further distinguished by variations of these characteristics, particularly those involving reproductive stages.

Annotated classification. The following classification is adapted from G.W. Martin in Ainsworth's *Dictionary of the Fungi,* 5th ed. (1961).

KINGDOM MYCOTA (fungi)

Eukaryotic (with true nuclei), achlorophyllous (without chlorophyll), acellular, unicellular, or multicellular organisms; microscopic or macroscopic in size; usually with cell walls and filaments; typically reproducing by spores produced asexually or sexually; walls containing chitin, cellulose, or both, among other substances; about 50,000 living species; fewer than 500 fossil species known.

Division Eumycota (true fungi)

Assimilative stage walled, typically filamentous (a mycelium), sometimes unicellular, usually eucarpic (having only part of the thallus forming a fruiting structure); asexual reproduction by fission, budding, fragmentation, or, more typically, by spores; sexual reproduction by various means, usually resulting in the formation of resting structures or meiospores.

Class Chytridiomycetes

Unicellular or filamentous, holocarpic (having all of the thallus involved in the formation of the fruiting body) or eucarpic; motile cells (zoospores or planogametes) characterized by a single, posterior, whiplash flagellum; mostly aquatic fungi saprobic or parasitic on algae, fungi, or, less often, on flowering plants.

Order Chytridiales. Mycelium lacking but rhizoids (short absorbing filaments) or rhizomycelium often present; chiefly freshwater saprobes or parasites of algae and fungi; some terrestrial species, such as *Olpidium brassicae* and *Synchytrium endobioticum*, cause plant disease; about 550 species.

Order Blastocladiales. Water molds with a restricted thallus, characterized by the production of thick-walled, pitted, resistant sporangia; sexual reproduction by isogamous (equal in size and alike in form) or anisogamous (unequal in size but still similar in form) planogametes; *Allomyces* exhibits an alternation of 2 equal generations; most are saprobes, but various species of *Coelomomyces* are parasitic in mosquito larvae; uniquely, their hyphae are devoid of cell walls; more than 50 species.

Order Monoblepharidales. Water molds with an extensive, foamy mycelium; sexual reproduction by a motile male gamete (antherozoid) fertilizing a nonmotile differentiated egg, resulting in a thick-walled oospore; about 20 species.

Class Hyphochytridiomycetes

A small group of mostly marine fungi very similar to the order Chytridiales but with motile cells bearing a single tinsel flagellum (*i.e.*, a flagellum with short side branches along the central axis, comblike).

Order Hyphochytriales. Characters of the class; about 15 species.

Class Plasmodiophoromycetes

Endoparasites (internal parasites) of fungi or plants often causing hypertrophy (excessive abnormal growth); assimilative stage an endophytic (living within plant tissue) plasmodium that becomes converted into a group of zoosporangia (structures producing motile asexual spores) or a large number of small, walled spores; motile cells with 2 unequal, anterior, whiplash flagella.

Order Plasmodiophorales. Characters of the class; *Plasmodiophora* and *Spongospora* cause serious plant diseases; about 35 species.

Class Oomycetes

Aquatic, amphibious, or terrestrial fungi; saprobic, facultatively (occasionally) or obligately (invariably) parasitic on plants, a few on fish; asexual reproduction typically by zoospores with 2 anterior or lateral flagella, 1 whiplash, 1 tinsel; sexual reproduction usually by contact of differentiated gametangia (gamete- or sex-cell-producing structures) with nuclei from the male fertilizing differentiated eggs and resulting in thick-walled oospores; thallus probably diploid with meiosis occurring in the gametangia.

Order Lagenidiales. Holocarpic, unicellular or filamentous water molds, parasitic on algae and fungi or saprobic; oogonium (egg-producing structure) typically containing a single egg; about 85 species.

Order Saprolegniales. Mostly eucarpic, filamentous water molds or soil fungi; saprobic or parasitic; hyphae without constrictions or cellulin plugs; oogonia containing 1 to many free eggs; some species are diplanetic, *i.e.*, they produce 2 types of zoospores, primary (pear-shaped with anterior flagella) and secondary (kidney-shaped with lateral flagella); some (*Aphanomyces*) cause root rots; others (*Saprolegnia*) infect fish and fish eggs; about 200 species.

Order Leptomitales. Aquatic saprobes found often in polluted waters; eucarpic; hyphae constricted, with cellulin plugs, arising from a well-defined basal cell; oogonium typically containing a single egg, which may be free or embedded in periplasm (a peripheral layer of protoplasm); 20 species.

Order Peronosporales. Aquatic or terrestrial; parasitic on algae or vascular plants, the latter mostly obligate parasites causing downy mildews; zoosporangia, in advanced species, borne on well-differentiated sporangiophores, deciduous and behaving as conidia (asexually produced spores); about 250 species.

Class Zygomycetes

Terrestrial saprobes or parasites of plants, animals, or humans; asexual reproduction by aplanospores (nonmotile spores) in sporangia or by conidia; sexual reproduction by fusion of morphologically similar gametangia, sometimes differing in size, resulting in thick-walled zygospores.

Order Mucorales. Often called the bread molds; saprobic, weakly parasitic on plants, or parasitic on humans and then causing mucormycosis (a pulmonary infection); asexual reproduction by sporangiospores, 1-spored sporangiola (a small deciduous sporangium), or conidia; in the genus *Pilobolus* the heavily cutinized sporangium is forcibly discharged; about 360 species.

Order Entomophthorales. Insect parasites or saprobes, some implicated in animal or human diseases; asexual reproduction by modified sporangia functioning as conidia, forcibly discharged; about 150 species.

Order Zoopagales. Parasitic on amoebas, rotifers, nematodes, or other small animals, which they trap by various specialized mechanisms; asexual reproduction by conidia borne singly or in chains, not forcibly discharged; about 60 species.

Class Trichomycetes

Commensals (organisms living parasitically on another organism but conferring some benefit in return, or at least not harming the host) with a filamentous thallus attached by a holdfast or basal cell to the digestive tract or external cuticle of living arthropods; asexual reproduction by sporangiospores (a spore borne within a sporangium), trichospores (zoospores or ciliated spores), arthrospores (a spore resulting from fragmentation of a hypha), or amoeboid cells; sexual reproduction, where known, zygomycetous.

Order Amoebidiales. Thallus coenocytic (without cross walls, with numerous freely distributed nuclei) arising from a holdfast; amoeboid cells formed; about 12 species.

Order Eccrinales. Thallus coenocytic, attached by a holdfast to the digestive tract of arthropods; aplanosporangia produced in succession; more than 50 species.

Order Asellariales. Thallus branched, septate, attached by a basal coenocytic cell; asexual reproduction by arthrospores; 6 species.

Order Harpellales. Thallus simple or branched, septate; asexual reproduction by trichospores; sexual reproduction zygomycetous; about 35 species.

Class Ascomycetes

Saprobic or parasitic on plants, animals, or humans; some are unicellular but most are filamentous, the hyphae septate with 1, rarely more, perforations in the septa; cells uninucleate or multinucleate; asexual reproduction by fission, budding, fragmentation, or, more typically, by conidia usually produced on special sporiferous (spore-producing) hyphae, the conidiophores, which are borne loosely on somatic (main-body) hyphae or variously assembled in asexual fruiting bodies; sexual reproduction by various means resulting in the production of meiosphores (ascospores) formed by free-cell formation in saclike structures (asci), which are produced naked or, more typically, are assembled in characteristic open or closed fruiting bodies (ascocarps); among the largest and most commonly known ascomycetes are the morels, cup fungi, saddle fungi, and truffles.

Subclass Hemiascomycetidae. Asci naked, formed from single cells or on hyphae; no ascocarps or ascogenous hyphae produced; saprobic or parasitic.

Order Protomycetales. Spore sac compound (a synascus); a poorly known small group of plant-parasitic ascomycetes; 20 or more species.

Order Endomycetales. Mostly saprobic, a few parasitic; zygote or single cell transformed directly into the ascus; mycelium sometimes lacking; this group includes the yeasts and their relatives.

Order Taphrinales. Parasites on vascular plants; asci produced from binucleate ascogenous (ascus-producing) cells formed from the hyphae in the manner of chlamydospores (thick-walled spores); 90 or more species.

Subclass Euascomycetidae. Asci unitunicate (ascus wall single-layered), borne in various types of ascocarps; saprobic or parasitic on plants, animals, or humans.

Order Eurotiales. Asci globose to broadly oval, typically borne at different levels in cleistothecia (completely closed ascocarp or fruiting structure); most of the human and animal dermatophytes belong here, also many saprobic soil or coprophilous fungi; possibly up to 150 species.

Order Microascales. Asci evanescent (quickly deteriorating), borne at different levels in perithecia (closed ascocarps with a pore in the top) with ostioles (the opening of the perithecium), or sometimes a long necklike structure terminating in a pore; some serious plant parasites such as *Ceratocystis ulmi* (Dutch elm disease) and *C. fagacearum* (oak wilt) belong here; about 100 species.

Order Onygenales. Asci formed in a mazaedium (a fruiting body consisting of a powdery mass of free spores interspersed with sterile threads, enclosed in a peridium or wall structure), evanescent, and liberating the ascospores as a powdery mass among sterile threads; about 25 species.

Order Erysiphales. Obligate parasites on flowering plants causing powdery mildews; mycelium white, superficial in most, feeding by means of haustoria sunken into the epidermal cells of the host; 1 to several asci in a cleistothecium, if more than 1, in a basal layer at maturity; asci globose to broadly oval; cleistothecia with appendages; about 150 species.

Order Meliolales. Mycelium dark, superficial on leaves and stems of vascular plants, typically bearing appendages (termed hyphopodia or setae); asci in basal layers in ostiolate perithecia without appendages; mostly tropical fungi; more than 1,000 species.

Order Chaetomiales. Asci in basal layers in superficial perithecia that bear conspicuous, straight or curly, simple or branched hairs on the surface; asci evanescent; about 110 species.

Order Xylariales. Perithecia with dark, membranous or carbonous (appearing as black burned wood) walls, with or without a stroma (a compact structure on or in which fructifications are formed); asci persistent, borne in a basal layer among paraphyses (elongate structures resembling asci but sterile), which may ultimately gelatinize and disappear; a rather large group of fungi one of which, *Neurospora crassa,* has been used extensively in genetic and biochemical studies; approximately 4,500 species.

Order Diaporthales. Perithecia immersed in plant tissue or in a stroma with their long ostioles protruding; ascal stalks gelatinizing, freeing the asci from their basal attachment; paraphyses lacking; the chestnut blight fungus (*Endothia parasitica*) belongs here; close to 500 species.

Order Hypocreales. Perithecia and stromata when present, brightly coloured, soft, fleshy, or waxy, when fresh; asci borne in a basal layer among apical paraphyses; about 800 species.

Order Clavicipitales. Perithecia immersed in a stroma that issues from a sclerotium (a hard-resting body resistant to unfavourable environmental conditions); asci with a thick apex penetrated by a central canal through which the septate, threadlike ascospores are ejected; the ergot fungus (*Claviceps purpurea*), cause of ergotism in plants, animals, and humans, and the original source of LSD, belongs to this order; some other Clavicipitales parasitize insect larvae; about 170 species.

Order Coryneliales. Asci in ascostromata with funnel-shaped ostioles at maturity; about 20 species.

Order Coronophorales. Asci in ascostromata with irregular or round, never funnel-shaped, openings; about 30 species.

Order Laboulbeniales. Ascomycetes of uncertain affinity; minute parasites of insects and arachnids with mycelium represented only by haustoria and stalks; about 1,635 species.

Order Ostropales. Ascocarp a loculelike apothecium (an open, often cuplike ascocarp); asci inoperculate (without a terminal pore) constructed as in the Clavicipitales; ascospores septate, threadlike; about 80 species.

Order Phacidiales. Ascocarp an apothecium immersed in a black stroma, the upper covering of which splits in stellate (star-shaped) or irregular fashion when ascospores mature; about 150 species.

Order Helotiales. Ascocarp an apothecium bearing inoperculate asci exposed from an early stage; some important plant diseases are caused by members of this group (for example, *Monilinia fructicola* causes brown rot of stone fruits), the earth tongues (*Geoglossaceae*) also belong here; more than 1,500 species.

Order Pezizales. Ascocarp an apothecium bearing operculate (with a hinged cap) asci above the ground; apothecia often large, cup- or saucer-shaped, spongy, brainlike, saddle-shaped, etc.; this group includes the morels, the false morels, and the saddle fungi among others; about 700 species.

Order Tuberales (truffles). The ascocarps, mostly closed and borne below the ground, are considered to be modified apothecia; the asci are globose, broadly oval, or club-shaped; about 230 species.

Subclass Loculoascomycetidae. Asci bitunicate, borne in ascostromata; saprobic or parasitic on plants.

Order Myriangiales. Asci borne singly in locules arranged at various levels in a more or less globose stroma; about 100 species.

Order Dothideales. Asci borne in fascicles (clusters) in a locule devoid of sterile elements; about 600 species.

Order Pleosporales. Asci borne in a basal layer among pseudoparaphyses; more than 2,000 species.

Order Microthyriales. Stroma flattened, hemispherical, opening by a pore or tear; base usually lacking; asci borne among pseudoparaphyses; mostly tropical fungi; about 1,200 species.

Order Hysteriales. Stroma boat-shaped, opening by a longitudinal slit, which renders it apothecium-like; asci borne among pseudoparaphyses; about 110 species.

Class Basidiomycetes

Saprobic or parasitic on plants or insects; filamentous; the hyphae septate, the septa typically inflated (dolipore) and centrally perforated; mycelium of 2 types, primary of uninucleate cells, succeeded by secondary, consisting of dikaryotic cells, this often bearing bridgelike clamp connections over the septa; asexual reproduction by fragmentation, oidia (thin-walled, free, hyphal cells behaving as spores), or conidia; sexual reproduction by fusion of hyphae (somatogamy), fusion of an oidium with a hypha (oidization), or fusion of a spermatium (a nonmotile male structure that empties its contents into a receptive female structure during plasmogamy—a kind of gamete) with a specialized receptive hypha (spermatization), resulting in dikaryotic hyphae that eventually give rise to basidia, either singly on the hyphae or in variously shaped basidiocarps; meiospores (basidiospores) borne on basidia; in the rusts (Uredinales) and smuts (Ustilaginales), the dikaryotic hyphae produce teleutospores (thick-walled resting spores), which are a part of the basidial apparatus; this is a large class of fungi containing the rusts, smuts, jelly fungi, club fungi, coral and shelf fungi, mushrooms, puffballs, stinkhorns, and bird's-nest fungi.

Subclass Heterobasidiomycetidae. Basidia septate or deeply divided or arising from a teleutospore or cyst; basidiospores often germinating by repetition, budding, or production of conidia; includes the jelly fungi, the rusts, and the smuts.

Order Tremellales (jelly fungi). Fruiting bodies (basidiocarps) well-formed, appearing as inconspicuous horny crusts when dry but usually bright-coloured to black gelatinous masses after a rain; a few are parasitic on mosses, vascular plants, or insects; most are saprobes; about 500 species.

Order Uredinales (rusts). Parasitic on vascular plants; basidial apparatus consists of a thick-walled teleutospore (probasidium), which either gives rise to a 4-celled tube (metabasidium) on which basidiospores are borne or which itself becomes 4-celled and produces basidiospores directly; basidiospores forcibly discharged; many rusts are heteroecious, *i.e.,* they require 2 species of host to complete their life cycle; rusts are among the fungi most destructive to agriculture; about 4,600 species.

Order Ustilaginales (smuts). Called smuts because the masses of spores (sori) are usually black and dusty; basidial apparatus consisting of a thick-walled teleutospore (probasidium), which, upon germination, gives rise to a septate or nonseptate tube (metabasidium), which bears the basidiospores; basidiospores not forcibly discharged, germinating usually by budding or by fusing and then producing a mycelial germ tube; various cereal smuts are of great economic importance; about 700 species.

Subclass Homobasidiomycetidae. Includes the great majority of the Basidiomycetes; most produce conspicuous, large-fruiting bodies, which bear the spores on basidia; basidia are simple, cylindrical, or club-shaped; basidiospores, which may or may not be forcibly discharged, germinate directly into a mycelium.

Order Exobasidiales. Basidiocarps lacking; basidia produced in a layer on the surface of parasitized vascular plants; 15 species.

Order Polyporales. Basidiocarps present; a large and probably heterogeneous order of fungi in which the basidia are borne in various ways but rarely on gills; includes the coral fungi, the club fungi, the chanterelles, and the pore (shelf or bracket) fungi among others; common genera include *Stereum, Clavaria, Hydnum, Cantharellus, Polyporus, Fomes; Schizophyllum* has been used extensively for genetic research; up to 2,500 species.

Order Agaricales (mushrooms and boletes). Basidia produced in layers (hymenia) on the underside of fleshy fruiting bodies (basidiocarps), in tubes (boletes) or on gills (mushrooms); some of these fungi form mycorrhizae, some are parasitic and cause root rots; most are saprobic; 4,000 to 5,000 species.

Order Hymenogastrales. Basidiocarps underground or on the surface but usually buried in humus, remaining closed, the interior (gleba) disintegrating into a slimy mass containing the spores; about 225 species.

Order Lycoperdales (puffballs). Gleba dry and powdery at maturity; consisting of small, pale spores and well-developed capillitium; about 160 species.

Order Sclerodermatales. These are puffballs with a hard peridium enclosing a dry, powdery gleba consisting of large, dark spores and some capillitium; about 120 species.

Order Phallales (stinkhorns). Gleba slimy and fetid at maturity; exposed on an elongated or net-shaped receptacle; *Phallus, Mutinus, Dictyophora, Simblum, Clathrus* are temperate-zone genera; about 70 species.

Order Nidulariales (bird's-nest fungi). The gleba separates into chambers, which become thick-walled, waxy, and hard—these are the peridioles ("eggs"), which are evident within a cuplike or gobletlike basidiocarp, the whole resembling a bird's nest at maturity—*Cyathus* and *Crucibulum* are the 2 most widely distributed genera; about 60 species.

Form-class Deuteromycetes

Fungi with septate mycelium reproducing only asexually and resembling asexual stages of Ascomycetes and Basidiomycetes.

Form-order Sphaeropsidales. Conidia borne in pycnidia; about 5,500 species.

Form-order Melanconiales. Conidia borne in acervuli; about 1,000 species.

Form-order Moniliales. Conidia borne on variously assembled conidiophores but never in pycnidia or acervuli; 10,000 or more species.

Form-order Mycelia Sterilia. No conidia produced; probably mycelial stages of Basidiomycetes; about 200 species.

Critical appraisal. Although the fungi were traditionally classified in the plant kingdom, some of their characteristics argue strongly against such affinities. Chief among them are their heterotrophy and the almost universal presence of chitin, which is also found in the skin of many invertebrate animals such as insects, in their walls. The classification followed in this article and in other articles in the *Ma-*

cropædia removes the fungi from the plant kingdom and includes them instead in a separate kingdom—Fungi. The slime molds (Myxomycetes, or Mycetozoa) are placed in the kingdom Protista. Older classification systems applied the name Phycomycetes (algal fungi) to the six classes of "lower fungi" (Chytridiomycetes, Hyphochytridiomycetes, Plasmodiophoromycetes, Oomycetes, Zygomycetes, and Trichomycetes). Although no longer recognized as a formal taxonomic category, the name Phycomycetes is still a useful term for designating those fungi that produce their spores in sporangia and have, in most cases, coenocytic hyphae. These "lower fungi" have been included by some protistologists in the kingdom Protista (see PROTISTS). A definitive classification of the "lower fungi" and protists remains to be agreed upon, however, and the classifications presented above and in the article PROTISTS reflect differences among scientists concerning taxonomy.

The groups that are designated as the form-class Deuteromycetes (also sometimes called "Fungi Imperfecti") consist of organisms whose sexual stages either have not been found or do not exist, together with the asexual stages of some sexually reproducing species. It has happened that sexual stages have been found for some species in this group, and, whenever this occurs, the species is immediately given a name appropriate to its proper taxonomic group. Sometimes, however, it has been found advisable to maintain the form-category name even though sexual stages are known, and this has led to the unusual condition of one species having been assigned two scientific names, only one of which, of course, is valid—the one applying to the sexual stage. The practice is justified mainly for the convenience of having a classification system, artificial though it is, into which conidial (asexual) forms may be grouped and studied. It has even been found useful to extend the concept of form categories to include the conidial stages of known sexually reproducing fungi, mainly Ascomycetes. Since many of these fungi, particularly the parasitic ones, are usually encountered only in the conidial stage, they can be easily identified by making use of the form-category classification. The convenience is great and the practice so widespread that in 1950 the International Botanical Congress legalized the use of form-names for conidial stages, recognizing, of course, the name of the perfect (sexual) stage as the official name of the whole organism. (C.J.Al./Ed.)

BIBLIOGRAPHY

General works: D.L. HAWKSWORTH, B.C. SUTTON, and G.C. AINSWORTH, *Ainsworth & Bisby's Dictionary of the Fungi*, 7th ed. (1983), remains the standard reference for terminology and definitions; WALTER H. SNELL and ESTHER A. DICK, *A Glossary of Mycology*, rev. ed. (1971), is an excellent dictionary of mycological terms; JOHN RAMSBOTTOM, *Mushrooms & Toadstools* (1953), offers a beautifully illustrated discussion of the occurrence and activities of one group of fungi; and CONSTANTINE J. ALEXOPOULOS and CHARLES W. MIMS, *Introductory Mycology*, 3rd ed. (1979), is an excellent text for both beginning and advanced students. ELIZABETH MOORE-LANDECKER, *Fundamentals of the Fungi*, 3rd ed. (1990), is a good introduction. Other introductions include LILIAN E. HAWKER, *Fungi*, 2nd ed. (1974); JOHN WEBSTER, *Introduction to Fungi*, 2nd ed. (1980); and J.H. BURNETT, *Fundamentals of Mycology*, 2nd ed. (1976), somewhat difficult for the novice. HAROLD C. BOLD, CONSTANTINE J. ALEXOPOULOS, and THEODORE DELEVORYAS, *Morphology of Plants and Fungi*, 5th ed. (1987); and C.T. INGOLD, *The Biology of Fungi*, 5th ed. (1984), offer surveys of the subject. FREDERICK A. WOLF and FREDERICK T. WOLF, *The Fungi*, 2 vol. (1947, reissued 1969), discusses morphology, taxonomy, physiology, genetics, ecology, and medical and industrial mycology. ERNST ATHEARN BESSEY, *Morphology and Taxonomy of Fungi* (1950, reprinted 1985), is a reference strong on phylogeny and the bibliography of classification. WILLIAM D. GRAY, *The Relation of Fungi to Human Affairs* (1959), discusses useful and destructive fungi, with emphasis on the application of mycology to industry; and D.L. HAWKSWORTH and B.E. KIRSOP (eds.), *Filamentous Fungi* (1988), explores applications of these fungi to biotechnology. G.C. AINSWORTH and ALFRED S. SUSSMAN (eds.), *The Fungi*, 4 vol. in 5 (1965–73), an advanced treatise written by specialists in various fields, discusses fungi in terms of cells, the organism, populations, and classification. A large number of topics on the biology of fungi are discussed by specialists in the following collections: JOHN E. SMITH, DAVID R. BERRY, and BJORN KRISTIANSEN (eds.), *The Filamentous Fungi*, 4 vol. (1975–83); ANTHONY H. ROSE and J. STUART HARRISON (eds.), *The Yeasts*, 2nd ed., 3 vol. (1987–89); and GARRY T. COLE and BRYCE KENDRICK (eds.), *Biology of Conidial Fungi*, 2 vol. (1981). Questions of morphological development are covered in JOHN E. SMITH (ed.), *Fungal Differentiation: A Contemporary Synthesis* (1983); PAUL J. SZANISZLO and JAMES L. HARRIS (eds.), *Fungal Dimorphism: With Emphasis on Fungi Pathogenic for Humans* (1985); and G. TURIAN and H.R. HOHL (eds.), *The Fungal Spore, Morphogenetic Controls* (1981). A.H. REGINALD BULLER, *Researches on Fungi*, 7 vol. (1909–50), is a classic collection of studies on various aspects of fungi, particularly strong on spore dispersal, development, and sexual reproduction; the first six volumes were reprinted in 1958. A useful guide of laboratory procedures for studying and handling fungi is presented in RUSSELL B. STEVENS (ed.), *Mycology Guidebook* (1974, reprinted with corrections and index by JOSEPH F. AMMIRATI, 1981). General discussion of physiological topics include MICHAEL O. GARRAWAY and ROBERT C. EVANS, *Fungal Nutrition and Physiology* (1984); DAVID H. GRIFFIN, *Fungal Physiology* (1981); and IAN K. ROSS, *Biology of the Fungi: Their Development, Regulation, and Associations* (1979). For an overview of progress in modern genetics of fungi, including genetic engineering, see J.W. BENNETT and LINDA L. LASURE (eds.), *Gene Manipulations in Fungi* (1985); WILLIAM E. TIMBERLAKE (ed.), *Molecular Genetics of Filamentous Fungi* (1985); and JOHN F. PEBERDY and LAJOS FERENCZY (eds.), *Fungal Protoplasts: Application in Biochemistry and Genetics* (1985).

Special subjects: LUCY KAVALER, *Mushrooms, Molds and Miracles* (1965), discusses various discoveries of fungal products. VALENTINA P. WASSON and R. GORDON WASSON, *Mushrooms, Russia, and History*, 2 vol. (1957), is an ethnomycological classic. C.T. INGOLD, *Dispersal in Fungi* (1953, reprinted with additions 1968), and *Spore Liberation* (1965), discuss the means by which fungi liberate their spores. KARL ESSER and RUDOLF KUENEN, *Genetics of Fungi* (1967; originally published in German, 1965), offers a general treatment; and JOHN R. RAPER, *Genetics of Sexuality in Higher Fungi* (1966), explores life cycles and sexual mechanisms in Ascomycetes and Basidiomycetes. D. PARKINSON and J.S. WAID (eds.), *The Ecology of Soil Fungi* (1960), is a series of essays; T.W. JOHNSON, JR., and F.K. SPARROW, JR., *Fungi in Oceans and Estuaries* (1961), provide important references; and WM. BRIDGE COOKE, *The Fungi of Our Mouldy Earth*, new ed. (1986), studies the fungi of the environment, with emphasis on water. CHESTER W. EMMONS *et al.*, *Medical Mycology*, 3rd ed. (1977); and CLYDE M. CHRISTENSEN, *Molds, Mushrooms, and Mycotoxins* (1975), study pathogenic fungi. J. WALTER WILSON and ORDA A. PLUNKETT, *The Fungous Diseases of Man* (1965), is a medical treatise. JOHN WILLARD RIPPON, *Medical Mycology: The Pathogenic Fungi and the Pathogenic Actinomycetes*, 3rd ed. (1988), is a textbook. E.C. LARGE, *The Advance of the Fungi* (1940, reprinted 1962), is a classic book on plant-disease fungi. Special biochemical topics are discussed in J.W. BENNETT and ALEX CIEGLER (eds.), *Secondary Metabolism and Differentiation in Fungi* (1983); and JOHN D. WEETE, *Lipid Biochemistry of Fungi and Other Organisms* (1980).

Special groups of fungi: CLYDE M. CHRISTENSEN, *Common Fleshy Fungi*, 2nd ed. (1955), is an easy-to-use manual for identifying common mushrooms; J. WALTON GROVES, *Edible and Poisonous Mushrooms of Canada*, rev. ed. (1979), identifies mushrooms; L.R. HESLER, *Mushrooms of the Great Smokies* (1960), is a field guide, with black-and-white photographs; RENÉ POMERLEAU and H.A.C. JACKSON, *Mushrooms of Eastern Canada and the United States*, trans. from French (1951), is a good manual; ALEXANDER H. SMITH, *Mushrooms in Their Natural Habitats*, 2 vol. (1949), is useful for the Pacific Northwest region; and GARY H. LINCOFF, *The Audubon Society Field Guide to North American Mushrooms* (1981), identifies many species. KENNETH B. RAPER and CHARLES THOM, *A Manual of the Penicillia* (1949, reprinted 1968); and KENNETH B. RAPER and DOROTHY I. FENNELL, *The Genus Aspergillus* (1965, reprinted 1977), are helpful. See also K.J. SCOTT and A.K. CHAKRAVORTY (eds.), *The Rust Fungi* (1982); and ROBERT W. LICHTWARDT, *The Trichomycetes, Fungal Associates of Arthropods* (1986).

Lichens: MASON E. HALE, JR., *The Biology of Lichens*, 3rd ed. (1983), is an intermediate-level introduction, and *How to Know the Lichens*, 2nd ed. (1979), is an authoritative guide. ANNIE L. SMITH, *Lichens* (1921, reprinted with additions 1975), a classic work on lichenology, is still a useful reference source. See also DAVID L. HAWKSWORTH and DAVID J. HILL, *The Lichen-Forming Fungi* (1984). YASUHIKO ASAHINA and SHOJI SHIBATA, *Chemistry of Lichen Substances* (1954, reprinted 1971; originally published in Japanese, 1949), is a classic reference book. BRUCE FINK, *The Lichen Flora of the United States* (1935, reissued 1971), is an advanced source. G.G. NEARING, *The Lichen Book* (1947, reissued 1962), is a good popular guide with drawings, popular names, and descriptions.

(C.J.Al./V.A./Ed.)

Galaxies

The stars of the universe are gathered together in numerous giant assemblages known as galaxies. Many such assemblages are so enormous that they contain hundreds of billions of stars. Nature has provided an immensely varied array of galaxies, ranging from faint, diffuse dwarf objects to brilliant, spiral-shaped giants. Virtually all galaxies appear to have been formed soon after the universe began, and they pervade space, even into the depths of the farthest reaches penetrated by powerful modern telescopes. Galaxies usually exist in clusters, some of which in turn are grouped into larger clusters measuring hundreds of millions of light-years across. (A light-year is the distance traversed by light in one year, traveling at a velocity of 300,000 kilometres per second, or 650,000,000 miles per hour.) These so-called superclusters are separated by nearly empty voids, causing the gross structure of the universe to look somewhat like a network of sheets and chains of galaxies.

Galaxies differ from one another in shape, with variations resulting from the way in which the systems were formed. Depending on the initial conditions in the pregalactic gas some 15,000,000,000 years ago, galaxies formed either as slowly turning, smoothly structured, round systems of stars and gas or as rapidly rotating pinwheels of such entities. Other differences between galaxies have been observed and are thought to reflect evolutionary changes. Some galaxies are rife with activity: they are the sites of star formation with its attendant glowing gas and clouds of dust and molecular complexes. Others, by contrast, are quiescent, having long ago ceased to form new stars. Perhaps the most conspicuous evolutionary changes in galaxies occur in their nuclei, where evidence suggests that in many cases supermassive objects—probably black holes—formed when the galaxies were young. Such phenomena occurred several billion years ago and are now observed as brilliant objects called quasars.

The existence of galaxies was not recognized until the early 20th century. Since then, however, galaxies have become one of the focal points of astronomical investigation. The notable developments and achievements in the study of galaxies are surveyed here. Much attention is also devoted to the Milky Way Galaxy—the local galaxy to which the Sun and Earth belong—and the galaxies and related objects that lie outside it. Included in the discussion are the properties, structures, and major components of the Milky Way system and the external galaxies; the distribution of the latter in clusters and superclusters; and the evolution of galaxies and quasars. For specifics about the components of galaxies, see STARS AND STAR CLUSTERS, NEBULA, and COSMOS.

For coverage of other related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 131 and 132, and the *Index*.

This article is divided into the following sections:

Historical survey of the study of galaxies

EARLY OBSERVATIONS AND CONCEPTIONS

The notion of spiral nebulas

The dispute over the nature of what were once termed spiral nebulas stands as one of the most significant in the development of astronomy. On this dispute hinged the question of the magnitude of the Cosmos: were we confined to a single, limited stellar system that lay embedded alone in empty space, or was our Milky Way Galaxy just one of millions of galaxies that pervaded space, stretching beyond the vast distances probed by our most powerful telescopes? How this question arose and how it was resolved is an important element in the development of our prevailing view of the universe.

Up until 1925 spiral nebulas and their related forms had uncertain status. Some scientists, notably Heber D. Curtis of the United States and Knut Lundmark of Sweden, had argued that they might be remote aggregates of stars similar in size to the Milky Way Galaxy. Centuries earlier the German philosopher Immanuel Kant, among others, had suggested much the same idea, but that was long before the tools were available to actually measure distances and thus prove it. During the early 1920s astronomers were divided. Although some deduced that spiral nebulas were actually extragalactic star systems, there was evidence that convinced many that such nebulas were local clouds of material, possibly new solar systems in the process of forming.

The problem of the Magellanic Clouds. It is now known that the nearest external galaxies are the Magellanic Clouds, two patchy, irregular objects visible in the skies of the Southern Hemisphere. For years, most experts who regarded the Clouds as portions of the Milky Way system separated from the main stream could not study them because of their position. (Both Magellanic Clouds are too far south to be seen from the United States.) Moreover, the irregular shapes of the objects and their numerous hot, blue stars, star clusters, and gas clouds did indeed make them resemble the southern Milky Way.

Discovery of the period–luminosity relation

The American astronomer Harlow Shapley, noted for his far-reaching work on the size and structure of the Milky Way Galaxy, was one of the first to appreciate the importance of the Magellanic Clouds in terms of the nature of spiral nebulas. To gauge the distance of the Clouds, he made use of the period–luminosity (P–L) relation discovered by Henrietta Leavitt of the Harvard College Observatory. In 1912 Leavitt had found that there was a close correlation between the periods of pulsation (variations in light) and the luminosities (intrinsic, or absolute, brightnesses) of a class of stars called Cepheid variables in the Small Magellanic Cloud. Leavitt's discovery, however, was of little practical value until Shapley worked out a calibration of the absolute brightnesses of pulsating stars closely analogous to the Cepheids, the so-called RR Lyrae variables (see below). With this quantified form of the P–L relation, he was able to calculate the distances to the Magellanic Clouds, determining that they were about 75,000 light-years from the Earth. The significance of the Clouds, however, continued to elude scientists of the time. For them, these objects still seemed to be anomalous, irregular patches of the Milky Way, farther away than initially thought but not sufficient to settle the question of the nature of the Cosmos.

Novas in the Andromeda Nebula. An unfortunate misidentification hampered the early recognition of the northern sky's brightest nearby galaxy, the Andromeda Nebula, also known as M31. In 1885 a bright star, previously invisible, appeared near the centre of M31, becoming almost bright enough to be seen without a telescope. As it slowly faded again, astronomers decided that it must be a nova, a "new star," similar to the class of temporary stars found relatively frequently in populous parts of the Milky Way. If this was the case, it was argued, then its extraordinary brightness must indicate that M31 cannot be very far away, certainly not outside of the local system of stars. Designated S Andromeda in conformity with the pattern of terminology applied to stars of variable brightness, this supposed nova was a strong argument in favour of the hypothesis that nebulas are nearby objects in the Milky Way system.

By 1910, however, there was evidence that S Andromeda might have been wrongly identified. Deep photographs were being taken of M31 with the Mount Wilson Observatory's newly-completed 152-centimetre (60-inch) telescope, and the astronomers at the observatory, especially J.C. Duncan and George W. Ritchey, were finding faint objects, just resolved by the longest exposures, which also seemed to behave like novas. These objects, however, were about 10,000 times fainter than S Andromeda. If they were ordinary novas then M31 must be millions of light-years away; but then the nature of S Andromeda became a difficult question. At this vast distance its total luminosity would have to be immense—an incomprehensible output of energy for a single star.

Completion of the 254-centimetre (100-inch) telescope on Mount Wilson in 1917 resulted in a new series of photographs that captured even fainter objects. More novas were found in M31, mainly by Milton L. Humason, who was an assistant at the time to Edwin P. Hubble, one of the truly outstanding astronomers of the day. Hubble eventually studicd 63 of these stars, and his findings proved to be one of the final solutions to the controversy (see below *The distance to the Andromeda Nebula*).

The scale of the Milky Way Galaxy. At the same time that spiral nebulas were bcing studied and debated, the Milky Way Galaxy became the subject of contentious discussion. During the early years of the 20th century, most astronomers believed that the Milky Way was a disk-shaped system of stars with the Sun near the centre and with the edge along a thick axis only about 15,000 light-years away. This view was based on statistical evidence involving star counts and the spatial distribution of a variety of cosmic objects—open star clusters, variable stars, binary systems, and clouds of interstellar gas. All of these objects seemed to thin out at distances of several thousand light-years.

Significance of Shapley's research on globular clusters

This conception of the Milky Way Galaxy was challenged by Shapley in 1917, when he released the findings of his study of globular clusters. He had found that these spherically symmetrical groups of densely packed stars, as compared to the much closer open clusters, were unusual in their distribution. While the known open clusters are concentrated heavily in the bright belt of the Milky Way, the globulars are for the most part absent from those areas, except in the general direction of the constellation Sagittarius where there is a concentration of faint globulars. Shapley's plot of the spatial distribution of these stellar groupings clarified this peculiar fact: the centre of the globular cluster system—a huge, almost spherical cloud of clusters—lies in that direction, some 30,000 light-years from the Sun. Shapley assumed that this centre must also be the centre of the Milky Way Galaxy. The globular clusters, he argued, form a giant skeleton around the disk of the Milky Way, and thus the system is immensely larger than was previously thought, its total extent measuring nearly 100,000 light-years (see below *The Milky Way Galaxy: Size and mass*).

Shapley succeeded in making the first reliable determination of the size of the Milky Way Galaxy largely by using Cepheids and RR Lyrae stars as distance indicators. His approach was based on the period–luminosity relation discovered by Leavitt and on the assumption that all of these variables have the same P–L relation. As he saw it, this assumption was most likely true in the case of the RR Lyrae stars, because all variables of this type in any given globular cluster have the same apparent brightness. If all RR Lyrae variables have the same intrinsic brightness, then it follows that differences in apparent brightness must be due to different distances from the Earth. The final step in developing a procedure for determining the distances of variables was to calculate the distances of a handful of such stars by an independent method so as to enable calibration. Shapley could not make use of the trigonometric parallax method since there are no variables close enough for direct distance measurement. However, he had recourse to a technique devised by the Danish astronomer Ejnar Hertzsprung that could determine distances to certain nearby field variables (*i.e.,* those not associated with any particular cluster) using measurements of their proper motions and the radial velocity of the Sun (see below *Stellar motions*). Accurate measurements of the proper motions of the variables based on long-term observations were available, and the Sun's radial velocity could be readily determined spectroscopically. Thus, by availing himself of this body of data and adopting Hertzsprung's method, Shapley was able to obtain a distance scale for Cepheids in the solar neighbourhood.

Shapley applied the zero point of the Cepheid distance scale to the globular clusters he had studied with the 152-centimetre telescope at Mount Wilson. Some of these clusters contained RR Lyrae variables, and for these Shapley could calculate distances in a straightforward manner from the period–luminosity relation. For other globulars he made distance determinations using a relationship that he discovered between the brightnesses of the RR Lyrae stars and the brightness of the brightest red stars. For still others he made use of apparent diameters, which he found to be relatively uniform for clusters of known distance. The final result was a catalog of distances for 69 globular clusters, from which Shapley deduced his revolutionary model of the Milky Way—one that not only significantly extended the limits of the galactic system but that also displaced the Sun from its centre to a location nearer its edge.

Shapley's work caused astronomers to ask themselves certain questions: How could the existing stellar data be so wrong? Why couldn't they see something in Sagittarius, the proposed galactic centre, 30,000 light-years away? The reason for the incorrectness of the star count methods was not learned until 1930, when Lick Observatory astronomer Robert J. Trumpler, while studying open clusters, discovered that interstellar dust pervades the plane of the Milky Way Galaxy and obscures objects beyond only a few thousand light-years. This dust thus renders the centre of the system invisible optically and makes it appear that globular clusters and spiral nebulas avoid the band of the Milky Way.

The zone of avoidance

Shapley's belief in the tremendous size of the local galactic system helped to put him on the wrong side of the argument about other galaxies. He thought that if the Milky Way was so immense, the spiral nebulas must lie within it. His conviction was reinforced by two lines of evidence. One of these has already been mentioned—the nova S Andromeda was so bright as to suggest that the Andromeda Nebula most certainly was only a few hundred light-years away. The second came about because of a very curious error made by one of Shapley's colleagues at Mount Wilson Observatory, Adrian van Maanen.

The van Maanen rotation. During the early 20th century one of the most important branches of astronomy was astrometry, the precise measurements of stellar positions and motions. Van Maanen was one of the leading experts in this field. Most of his determinations of stellar positions were accurate and have stood the test of time, but he made one serious and still poorly understood error when he pursued a problem tangential to his main interests. In a series of papers published in the early 1920s, van Maanen reported on his discovery and measurement of the rotation of spiral nebulas. Using early 152-centimetre plates taken by others as well as more recent ones taken about 10 years later, van Maanen measured the positions of several knotlike, nearly stellar images in the spiral arms of some of the largest known spiral nebulas (*e.g.*, M33, M101, and M51). Comparing the positions, he found distinct changes indicative of a rotation of the spiral pattern against the background of surrounding field stars. In each case, the rotation occurs in the sense that the spiral arms trail. The periods of rotation were all approximately 100,000 years. Angular motions were about 0.02 second of arc per year.

Shapley seized the van Maanen results as evidence that the spirals must be nearby; otherwise, their true space velocities of rotation would have to be impossibly large. For example, if M51 is rotating at an apparent rate of 0.02 second of arc per year, its true velocity would be immense if it is a distant galaxy. Assuming that a distance of 10,000,000 light-years would lead to an implausibly large rotation velocity of 12,000 kilometres per second (km/sec), Shapley argued that if a more reasonable velocity were adopted, say, 100 km/sec, then the distances would be all less than 100,000 light-years, putting all the spirals well within the Milky Way Galaxy.

It is unclear just why such a crucial measurement went wrong. Van Maanen repeated the measures and obtained the same answer even after Hubble demonstrated the truth about the distances to the spirals (see below). However, subsequent workers, using the same plates, failed to find any rotation. Among the various hypotheses that science historians have proposed as an explanation for the error are two particularly reasonable ideas: (1) possibly the fact that spiral nebulas look like they are rotating (*i.e.*, they resemble familiar rotational patterns that are perceivable in nature) may have influenced the observer subconsciously, and this subtle effect manifested itself in prejudicing the delicate measurements; or (2) possibly the first set of plates was the problem. Many of these plates had been taken in an unconventional manner by Ritchey, who swung the plateholder out of the field whenever the quality of the images was temporarily poor because of atmospheric turbulence. The resulting plates appeared excellent, having been exposed only during times of very fine seeing; however, according to some interpretations, the images had a slight asymmetry that led to a very small displacement of star images compared to nonstellar images. Such an error could look like rotation if not recognized for what it really was. In any case, the van Maanen rotation was accepted by many astronomers, including Shapley, and temporarily sidetracked progress toward recognizing the truth about galaxies.

The Shapley–Curtis debate. The nature of galaxies and scale of the Cosmos were the subject of the "Great Debate," a public program arranged in 1920 by the National Academy of Sciences at the Smithsonian Institution in Washington, D.C. Featured were talks by Shapley and the aforementioned Heber Curtis, who were recognized as spokesmen for opposite views on the nature of spiral nebulas and the Milky Way. This so-called debate has often been cited as an illustration of how revolutionary new concepts are assimilated by science. It is sometimes compared to the debate, centuries before, over the motions of the Earth (the Copernican revolution); and, though as a focal point it can be used to define the modern controversy, it actually was much more complicated.

A careful reading of the documents involved suggests that on the broader topic of the scale of the universe, both men were making incorrect conclusions but for the same reasons—namely, for being unable to accept and comprehend the incredibly large scale of things. Shapley correctly argued for an enormous Milky Way Galaxy on the basis of the period–luminosity relation and the globular clusters, while Curtis incorrectly rejected these lines of evidence, advocating instead a small galactic system. Given a Milky Way system of limited scale, Curtis could argue for and consider plausible the extragalactic nature of the spiral nebulas. Shapley, on the other hand, incorrectly rejected the island universe theory of the spirals (*i.e.*, the hypothesis that there existed beyond the boundaries of the Milky Way comparable galaxies) because he felt that such objects would surely be engulfed by the local galactic system. Furthermore, he put aside the apparent faint novas in M31, preferring to interpret S Andromeda as an ordinary nova, for otherwise that object would have been unbelievably luminous. Unfortunately for him such phenomena—called supernovas—do in fact exist, as was realized a few years later. Curtis was willing to concede that there might be two classes of novas; yet, because he considered the Milky Way to be small, he underestimated their differences. The van Maanen rotation also entered into Shapley's arguments: if spiral nebulas were rotating so fast, they must be within the Milky Way as he conceived it. For Curtis, however, the matter provided less of a problem: even if spiral nebulas did rotate as rapidly as claimed, the small scale of Curtis' universe allowed them to have physically reasonable speeds.

Shapley's rejection of the island universe theory

The Shapley–Curtis debate took place near the end of the era of the single-galaxy universe. In just a few years the scientific world became convinced that Shapley's grand scale of the Milky Way Galaxy was correct and at the same time that Curtis was right about the nature of spiral nebulas. Such objects indeed lie even outside Shapley's enormous Milky Way, and they range far beyond the distances that in 1920 seemed too vast for many astronomers to comprehend.

Hubble's discovery of extragalactic objects. During the early 1920s Hubble detected 15 stars in the small, irregular cloudlike object NGC 6822 that varied in luminosity, and he suspected that they might include Cepheids. After considerable effort, he determined that 11 of them were in fact Cepheid variables, with properties indistinguishable from those of normal Cepheids in the Milky Way Galaxy and in the Magellanic Clouds. Their periods ranged from 12 to 64 days, and they were all very faint, much fainter than their Magellanic counterparts. Nevertheless, they fit a period–luminosity relation of the same nature as had been discovered by Leavitt (see above).

Hubble then boldly assumed that the P–L relation was universal and derived an estimate for the distance to NGC 6822 using Shapley's most recent (1925) version of the calibration of the relation. This calibration was wrong, as is now known, because of the confusion at that time over the nature of Cepheids. Shapley's calibration included certain Cepheids in globular clusters that subsequent investigators found to have their own fainter P–L relation. (Such Cepheids have been designated Type II Cepheids to distinguish them from the normal variety, which are referred to as Type I.) Thus, Hubble's distance for NGC 6822

NGC 6822 as the first recognized external galaxy

was too small: he calculated a distance of only 700,000 light-years. Today it is recognized that the actual distance is closer to 2,000,000 light-years. In any case, this vast distance—even though underestimated—was large enough to convince Hubble that NGC 6822 must be a remote, separate galaxy, much too far away to be included even in Shapley's version of the Milky Way system. Technically, then, this faint nebula can be considered to be the first recognized external galaxy. The Magellanic Clouds continued to be regarded simply as appendages to the Milky Way, and the other bright nebulas, M31 and M33, were still being studied at the Mount Wilson Observatory. Although Hubble did announce his discovery of Cepheids in M31 at a meeting in 1924, he did not complete his research and publish the results for this conspicuous spiral galaxy until five years later.

While the Cepheids made it possible to determine the distance and nature of NGC 6822, some of its other features corroborated the conclusion that it was a separate, distant galaxy. Hubble discovered within it five diffuse nebulas, which are glowing gaseous clouds composed mostly of ionized hydrogen designated H II regions. (H stands for hydrogen and II indicates that most of it is ionized; H I, by contrast, signifies neutral hydrogen.) He found that these five H II regions had spectra like those of gas clouds in the Milky Way system—*e.g.,* the Orion and Eta Carinae nebulas. Calculating their diameters, Hubble ascertained that the sizes of the diffuse nebulas were normal, similar to those of local examples of giant H II regions.

Five other diffuse objects discerned by Hubble were definitely not gaseous nebulas. He compared them with globular clusters (both in the Milky Way Galaxy and in the Magellanic Clouds) and concluded that they were too small and faint to be normal globular clusters. Convinced that they were most likely distant galaxies seen through NGC 6822, he dismissed them from further consideration. Modern studies suggest that Hubble was too hasty. Though probably not true giant globular clusters, these objects are in all likelihood star clusters in the system, fainter, smaller in population, and probably somewhat younger than normal globular clusters.

The uniformity of nature

The Dutch astronomer Jacobus Cornelius Kapteyn showed in the early 20th century that statistical techniques could be used to determine the stellar luminosity function for the solar neighbourhood. (The luminosity function is a curve that shows how many stars there are in a given volume for each different stellar luminosity.) Anxious to test the nature of NGC 6822, Hubble counted stars in the galaxy to various brightness limits and found a luminosity function for its brightest stars. When he compared it with Kapteyn's, the agreement was excellent—another indication that the Cepheids had given about the right distance and that the basic properties of galaxies were fairly uniform. Step by step, Hubble and his contemporaries piled up evidence for the fundamental assumption that has since guided the astronomy of the extragalactic universe, the uniformity of nature. By its bold application, astronomers have moved from a limiting one-galaxy Cosmos to an immense vastness of space populated by billions of galaxies, all grander in size and design than the Milky Way system had once been thought to be.

The distance to the Andromeda Nebula. In 1929 Hubble published his epochal paper on M31, the great Andromeda Nebula. Based on 350 photographic plates taken at Mount Wilson, his study provided evidence that M31 is a giant stellar system like the Milky Way Galaxy.

Because M31 is much larger than the field of view of the 152-centimetre and 254-centimetre telescopes at Mount Wilson, Hubble concentrated on four regions, centred on the nucleus and at various distances along the major axis. The total area studied amounted to less than half the galaxy's size, and the other unexplored regions remained largely unknown for 50 years. (Modern comprehensive optical studies of M31 have only been conducted since about 1980.)

Hubble pointed out an important and puzzling feature of the resolvability of M31. Its central regions, including the nucleus and diffuse nuclear bulge, were not well resolved into stars, one reason that the true nature of M31 had previously been elusive. However, the outer parts along the spiral arms in particular were resolved into swarms of faint stars, seen superimposed over a structured background of light. Current understanding of this fact is that spiral galaxies typically have central bulges made up exclusively of very old stars, the brightest of which are too faint to be visible on Hubble's plates. Only in 1944 did the German-born astronomer Walter Baade finally resolve the bulge of M31. Using red-sensitive plates and very long exposures, he managed to detect the brightest red giants of this old population. Out in the arms there exist many young, bright, hot, blue stars, and these are easily resolved. The brightest are so luminous that they can even be seen with moderate-size telescopes.

Cepheid variables in the Andromeda Galaxy

The most important of Hubble's discoveries was that of M31's population of Cepheid variables. Forty of the 50 variables detected turned out to be ordinary Cepheids with periods ranging from 10 to 48 days. A clear relation was found between their periods and luminosities, and the slope of the relation agreed with those for the Magellanic Clouds and NGC 6822. Hubble's comparison indicated that M31 must be 8.5 times more distant than the Small Magellanic Cloud (SMC), which would imply a distance of 2,000,000 light-years if the modern SMC distance were used (the 1929 value employed by Hubble was about two times too small). Clearly, M31 must be a distant, large galaxy.

Other features announced in Hubble's paper were M31's population of bright, irregular, slowly varying variables. One of the irregulars was exceedingly bright; it is among the most luminous stars in the galaxy and is a prototype of a class of high-luminosity stars now called Hubble–Sandage variables, which are found in many giant galaxies. Eighty-five novas, all behaving very much like those in the Milky Way Galaxy, were also analyzed. Hubble estimated that the true occurrence rate of novas in M31 must be about 30 per year, a figure that was later confirmed by the American astronomer Halton C. Arp in a systematic search.

Hubble found numerous star clusters in M31, especially globular clusters, 140 of which he eventually cataloged. He clinched the argument that M31 was a galaxy similar to the Milky Way by calculating its mass and mass density. Using the velocities that had been measured for the inner parts of M31 by spectrographic work, he calculated (on the basis of the distance derived from the Cepheids) that M31's mass must be about 3,500,000,000 times that of the Sun. Today, astronomers have much better data, which indicate that the galaxy's true total mass must be at least 100 times greater than Hubble's value, but even that value clearly showed that M31 is an immense system of stars. Furthermore, Hubble's estimates of star densities demonstrated that the stars in the outer arm areas of M31 are spread out with about the same density as in the Milky Way system in the vicinity of the Sun.

RECENT DEVELOPMENTS

Until about 1950 scientific knowledge of galaxies advanced slowly. Only a very small number of astronomers took up galaxy studies, and only a very few telescopes were suitable for significant research. It was an exclusive field, rather jealously guarded by its practitioners, and so progress was orderly but limited.

During the decade of the 1950s the field began to change. Ever larger optical telescopes became available, and the space program resulted in a sizable increase in the number of astronomers emerging from universities. New instrumentation enabled investigators to explore galaxies in entirely new ways, making it possible to detect their radio radiation, infrared and ultraviolet emissions, and eventually even radiation at X-ray wavelength. Whereas in the 1950s there was only one telescope larger than 254 centimetres (100 inches) and only about 10 astronomers conducting research on galaxies worldwide, by the 1980s the number of large telescopes had grown 10-fold and the number of scientists devoted to galaxy study was in the hundreds. Galaxies are now extensively studied with giant arrays of ground-based radio telescopes, Earth-orbiting X-ray, ultraviolet, and infrared telescopes, and high-speed electronic computers, giving rise to remarkable advances in knowledge and understanding.

Types of galaxies

PRINCIPAL SCHEMES OF CLASSIFICATION

Almost all current systems of galaxy classification are outgrowths of the initial scheme proposed by Hubble in 1926. In Hubble's scheme, which is based on the optical appearance of galaxy images on photographic plates, galaxies are divided into three general classes: ellipticals, spirals, and irregulars. His basic definitions are as follows:

Elliptical galaxies. Galaxies of this class have smoothly varying brightnesses, with the degree of brightness steadily decreasing outward from the centre. They appear elliptical in shape, with lines of equal brightness made up of concentric and similar ellipses. These galaxies are nearly all of the same colour: they are somewhat redder than the Sun.

Spiral galaxies. These galaxies are conspicuous for their spiral-shaped arms, which emanate from or near the nucleus and gradually wind outward to the edge. There are usually two opposing arms arranged symmetrically around the centre. The nucleus of a spiral galaxy is a sharp-peaked area of smooth texture, which can be quite small or, in some cases, can make up the bulk of the galaxy. The arms are embedded in a thin disk of stars. Both the arms and the disk of a spiral system are blue in colour, whereas its central areas are red like an elliptical galaxy.

Irregular galaxies. Most representatives of this class consist of grainy, highly irregular assemblages of luminous areas. They have no noticeable symmetry nor obvious central nucleus, and they are generally bluer in colour than are the arms and disks of spiral galaxies. An extremely small number of them, however, are red and have a smooth, though nonsymmetrical, shape.

Hubble subdivided these three classes into finer groups according to subtle differences in shape, as described in detail below. Other classification schemes similar to Hubble's follow this pattern but subdivide the galaxies differently. A notable example of one such system is that of Gerard de Vaucouleurs. This scheme, which has evolved considerably since its inception in 1959, includes a large number of codes for indicating different kinds of morphological characteristics visible in the images of galaxies. The major Hubble galaxy classes form the framework of de Vaucouleurs's scheme, and its subdivision includes different families, varieties, and stages, as shown in Table 1.

Examples of the de Vaucouleurs classification scheme are for galaxy M33, the Triangulum Nebula, which is classified as SA(s)cd, and the nearby small galaxy NGC 6822, classified as IB(s)m.

Van den Bergh's luminosity classification

An entirely different kind of classification scheme is the luminosity classification developed in 1960 by Sidney van den Bergh. Based on morphological considerations, luminosity classes are assigned to individual galaxies within the Hubble classes. Those that are the most luminous are given a luminosity class of I, and the intrinsically faintest members of a class are assigned a V or VI, recalling the general approach of the luminosity class scheme used for stellar spectra (see STARS AND STAR CLUSTERS). Thus a very luminous galaxy with open, resolved arms would be an Sc I galaxy, while a somewhat intrinsically fainter object with the same basic structure would be an Sc II or Sc III galaxy. To assign a luminosity class, a galaxy's image has to be compared with a set of standard images of galaxies for which distances are known and for which luminosity classes have been established by van den Bergh.

Classification schemes based on criteria other than optical appearance have been proposed. There is, for example, the Morgan scheme (proposed by W.W. Morgan), which combines information on the spectrum of a galaxy with its general shape. Here, a class is coded with a letter that indicates the spectral type of the galaxy in the blue (either as measured or as determined from the galaxy's bulge morphology, which correlates with the spectral type): *e.g.*, a, af, f, fg, g, gk, k, for increasing dominance by cooler stars. The code then includes a capital letter to indicate general morphology—*e.g.*, E, S, or I—in accordance with Hubble's general classes. This is followed by a number that indicates the overall optical shape of the image, with 0 representing a circular image and a 10 (never actually realized) standing for a linear, infinitely thin image. An example is the galaxy M31, the Andromeda Nebula, which is classified as kS5 in the Morgan system.

Table 1: De Vaucouleurs's Classification of Galaxies

classes	families	varieties	stages	type
Ellipticals				E
			elliptical (0–7)	E0
			intermediate	E0-1
			late elliptical	E$^+$
Lenticulars				S0
	ordinary			SA0
	barred			SB0
	mixed			SAB0
		inner ring		S(r)0
		S-shaped		S(s)0
		mixed		S(rs)0
			early	S0$^-$
			intermediate	S0°
			late	S0$^+$
Spirals	ordinary			SA
	barred			SB
	mixed			SAB
		inner ring		S(r)
		S-shaped		S(s)
		mixed		S(rs)
			0/a	S0/a
			a	Sa
			ab	Sab
			b	Sb
			bc	Sbc
			c	Sc
			cd	Scd
			d	Sd
			dm	Sdm
			m	Sm
Irregulars	ordinary			IA
	barred			IB
	mixed			IAB
		S-shaped		I(s)
			Magellanic	Im
			non-Magellanic	I0
Peculiars				P
Peculiarities (all types)			peculiarity	P
			uncertain	:
			doubtful	?
			spindle	sp
			outer ring	(R)
			pseudo outer ring	(R)

Classification of radio galaxies

Systems that separate galaxies according to the character of their radio structure and the strength of their radio emissions also have been devised. For example, radio galaxies can be classified according to the following scheme:

g: galaxies with normal radio fluxes.

R: galaxies with strong radio emission. Many have distorted morphology, with evidence of explosive events or interactions with companions.

cD: galaxies with abnormally large, distended shapes, always found in the central areas of galaxy clusters and hypothesized to consist of merged galaxies.

S: Seyfert galaxies, originally recognized by the American astronomer Carl K. Seyfert from optical spectra. These objects have very bright nuclei with strong emission lines of hydrogen and other common elements, showing velocities of hundreds or thousands of kilometres per second. Most are radio sources.

N: galaxies with small, very bright nuclei and strong radio emission, probably similar to Seyfert galaxies but more distant.

Q: quasars, small, extremely luminous objects, many of which are strong radio sources. Quasars apparently are related to Seyfert and N galaxies but have such bright nuclei that the underlying galaxy can be detected only with great difficulty.

Although such schemes are sometimes used for special purposes, including, for example, certain kinds of statistical studies, the general scheme of Hubble in its updated form is the one most commonly used and so will be described in detail in the following section.

CLASSES OF GALAXIES

In *The Hubble Atlas of Galaxies* (1961), Allan R. Sandage drew on Hubble's notes and his own research on galaxy morphology to revise the Hubble classification scheme. Some of the features of this revised scheme are subject to argument because of the findings of very recent research, but its general features, especially the coding of types, remain viable. A description of the classes as defined by Sandage is given here, along with observations concerning needed refinements of some of the details.

Elliptical galaxies. These systems exhibit certain characteristic properties. They have complete rotational symmetry; *i.e.*, they are figures of revolution with two equal principal axes. They have a third smaller axis that is the presumed axis of rotation. The surface brightness of ellipticals at optical wavelengths decreases monotonically outward from a maximum value at the centre, following a common mathematical law of the form:

$$I = I_o(r/a + 1)^{-2},$$

where I is the intensity of the light, I_o is the central intensity, r is the radius, and a is a scale factor. The isophotal contours exhibited by an elliptical system are similar ellipses with a common orientation, each centred on its nucleus. No galaxy of this type is flatter than $b/a = 0.3$, with b and a the minor and major axes of the elliptical image, respectively. Ellipticals contain neither interstellar dust nor bright stars of spectral types 0 and B. Many, however, contain evidence of the presence of low-density gas in their nuclear regions. Ellipticals are red in colour, and their spectra indicate that their light comes mostly from old stars, especially evolved red giants.

Subclasses of elliptical galaxies

Subclasses of elliptical galaxies are defined by their apparent shape, which is of course not necessarily their three-dimensional shape. The designation is E*n*, where *n* is an integer defined by

$$n = 10(a - b)/a.$$

A perfectly circular image will be an E0 galaxy, while a flatter object might be an E7 galaxy. (As explained above, elliptical galaxies are never flatter than this, so there are no E8, E9, or E10 galaxies.)

Although the above-cited criteria are generally accepted, current high-quality measurements have shown that some significant deviations exist. Most elliptical galaxies do not, for instance, exactly fit the intensity law formulated by Hubble; deviations are evident in their innermost parts and in their faint outer parts. Furthermore, many elliptical galaxies have slowly varying ellipticity, with the images being more circular in the central regions than in the outer parts. The major axes sometimes do not line up either, their position angles varying outward. Finally, astronomers have found that a few ellipticals do in fact have small numbers of luminous 0 and B stars as well as dust lanes.

Spiral galaxies. Spirals are characterized by circular symmetry, a bright nucleus surrounded by a thin outer disk, and a superimposed spiral structure. They are divided into two parallel classes: normal spirals and barred spirals. The normal spirals have arms that emanate from the nucleus, while barred spirals have a bright linear feature called a bar straddling the nucleus, with the arms unwinding from the ends of the bar. The normal spirals are designated "S" and the barred varieties "SB." Each of these classes is subclassified into three types according to the size of the nucleus and the degree to which the spiral arms are coiled. The three types are denoted with the lowercase letters "a," "b," and "c." There also exist galaxies that are intermediate between ellipticals and spirals. Such systems have the disk shape characteristic of the latter but no spiral arms. These intermediate forms bear the designation "S0" (Figure 1).

Normal and barred spirals

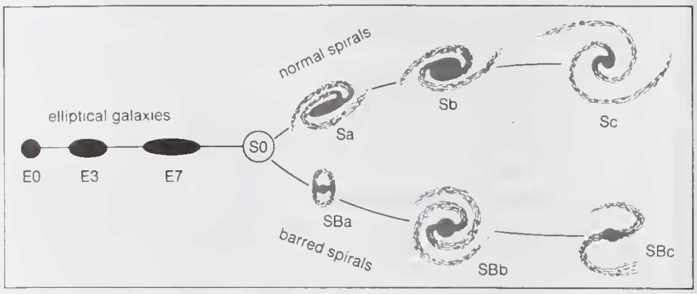


Figure 1: Hubble's system of classification for galaxies (see text).

S0 galaxies. These systems exhibit some of the properties of both the ellipticals and the spirals and seem to be a bridge between these two more common galaxy types. Hubble introduced the S0 class long after his original classification scheme had been universally adopted largely because he noticed the dearth of highly flattened objects that otherwise had the properties of elliptical galaxies. Sandage's elaboration of the S0 class yielded the characteristics described here.

S0 galaxies have a bright nucleus that is surrounded by a smooth, featureless bulge and a faint outer envelope. They are thin; statistical studies of the ratio of the apparent axes (seen projected onto the sky) indicate that they have intrinsic ratios of minor to major axes in the range 0.1 to 0.3. Their structure does not generally follow the luminosity law of elliptical galaxies, but it has a form more like that for spiral galaxies. Some S0 systems have a hint of structure in the envelope, either faintly discernible armlike discontinuities or narrow absorption lanes produced by interstellar dust. Several S0 galaxies are otherwise peculiar, and it is difficult to classify them with certainty. They can be thought of as peculiar Irr galaxies (*i.e.*, Irr II galaxies [see below]) or simply as some of the 1 or 2 percent of galaxies that do not fit easily into the Hubble scheme. Among these are such galaxies as NGC 4753 that has irregular dust lanes across its image and NGC 128 that has a double, almost rectangular, bulge around a central nucleus. Another type of peculiar S0 is found in NGC 2685. This nebula in the constellation Ursa Major has an apparently edge-on disk galaxy at its centre, with surrounding hoops of gas, dust, and stars arranged in a plane that is at right angles to the apparent plane of the central object.

Sa galaxies. These normal spirals have narrow, tightly wound arms, which usually are visible due to the presence of interstellar dust and in many cases bright stars as well. Most of them have a large, amorphous bulge in the centre, but there are some that violate this criterion, having a small nucleus around which is arranged an amorphous disk with superimposed faint arms. NGC 1302 is an example of the normal type of Sa galaxy, while NGC 4866 is representative of one with a small nucleus and arms consisting of thin dust lanes on a smooth disk.

Sb galaxies. This intermediate type of spiral typically has a medium-sized nucleus. Its arms are more widely spread than those of the Sa variety and appear less smooth. They contain stars, star clouds, and interstellar gas and dust. Sb galaxies show wide dispersions in details in terms of their shape. Hubble and Sandage observed, for example, that in certain Sb galaxies the arms emerge at the nucleus, which is often quite small. Other members of this subclass have arms that begin tangent to a bright, nearly circular ring, while still others reveal a small, bright spiral pattern inset into the nuclear bulge. In any of these cases, the spiral arms may be set at different pitch angles. (A pitch angle is defined as the angle between an arm and a circle centred on the nucleus and intersecting the arm.)

Hubble and Sandage noted further deviations from the standard shape established for Sb galaxies. A few systems exhibit a chaotic dust pattern superimposed upon the tightly wound spiral arms. Some have smooth, thick arms of low surface brightness, frequently bounded on their inner edges with dust lanes. Finally, there are those with a large, smooth nuclear bulge from which the arms emanate, flowing outward tangent to the bulge and forming short arm segments. This is the most familiar type of Sb galaxy and is best exemplified by the giant Andromeda Galaxy.

Many of these variations in shape remain unexplained. Theoretical models of spiral galaxies based on a number of different premises can reproduce the basic Sb galaxy

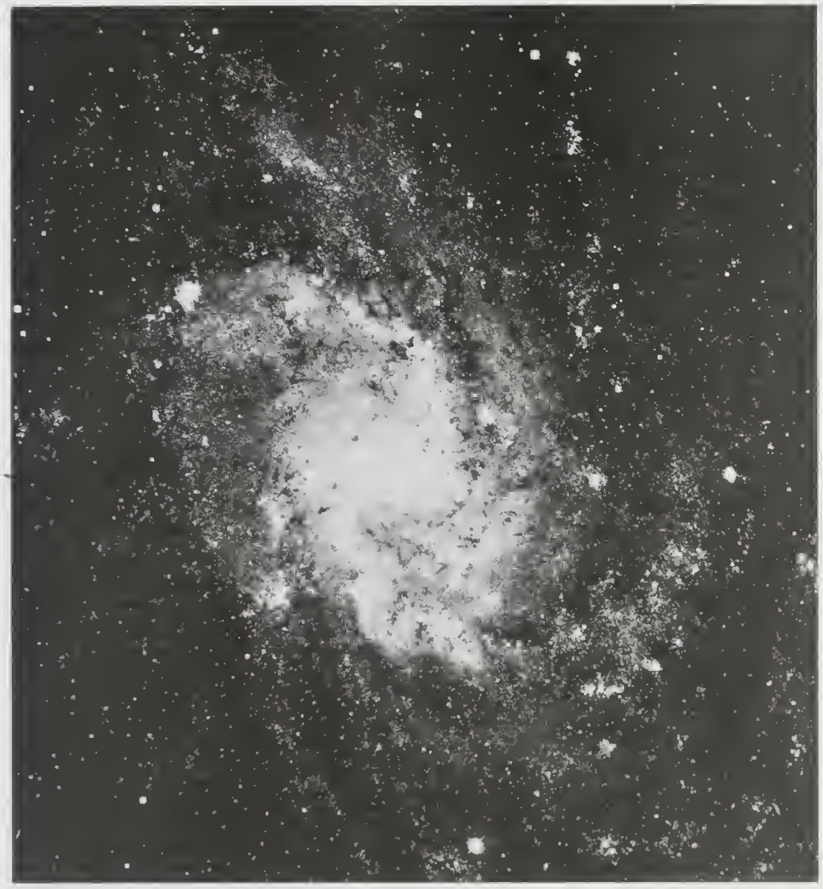


Figure 2: M33, a spiral galaxy in the constellation Triangulum.

By courtesy of the National Optical Astronomy Observatories

shape (see below *The Milky Way Galaxy*), but many of the deviations noted above are somewhat mysterious in origin and must await more detailed and realistic modeling of galactic dynamics.

Sc galaxies. These galaxies characteristically have a very small nucleus and multiple spiral arms that are open, with relatively large pitch angles. The arms, moreover, are lumpy, containing as they do numerous irregularly distributed star clouds, stellar associations, star clusters, and gas clouds known as emission nebulas.

As in the case of Sb galaxies, there are several recognizable subtypes among the Sc systems. Sandage has cited six subdivisions: (1) galaxies, such as the Whirlpool Nebula (M51), that have thin, branched arms that wind outward from a tiny nucleus, usually extending out about 180° before branching into multiple segments; (2) systems with multiple arms that start tangent to a bright ring centred on the nucleus; (3) those with arms that are poorly defined and that span the entire image of the galaxy; (4) those with a spiral pattern that cannot easily be traced and that are multiple and punctuated with chaotic dust lanes; (5) those with thick, loose arms that are not well defined—*e.g.*, the nearby galaxy M33, the Triangulum Nebula (Figure 2); and (6) transition types, which are almost so lacking in order that they could be considered irregular galaxies.

Some classification schemes, such as that of de Vaucouleurs, give the last of the above-cited subtypes a class of its own, type Sd (Figure 3). It also has been found that some of the variations noted here for Sc galaxies are related to total luminosity. Galaxies of the fifth subtype, in particular, tend to be intrinsically faint, while those of the first subtype are among the most luminous spirals known. This correlation is part of the justification for the luminosity classification discussed earlier.

SB galaxies. The luminosities, dimensions, spectra, and distributions of the barred spirals tend to be indistinguishable from those of normal spirals. The subclasses of SB systems exist in parallel sequence to those of the latter.

Subclasses of SB galaxies

There are SB0 galaxies that feature a large nuclear bulge surrounded by a disklike envelope across which runs a luminous, featureless bar. Some SB0 systems have short bars, while others have bars that extend across the entire visible image. Occasionally there is a ringlike feature external to the bar. SBa galaxies have bright, fairly large nuclear bulges and tightly wound, smooth spiral arms that emerge from the ends of the bar or from a circular ring external to the bar. SBb systems have a smooth bar as well as relatively smooth and continuous arms. In some galaxies of this type, the arms start at or near the ends of the bar, with conspicuous dust lanes along the inside of the bar that can be traced right up to the nucleus. Others have arms that start tangent to a ring external to the bar. In SBc galaxies, both the arms and the bar are highly resolved into star clouds and stellar associations. The arms are open in form and can start either at the ends of the bar or tangent to a ring.

From Paul W Hodge *The Physics and Astronomy of Galaxies and Cosmology* (1966)

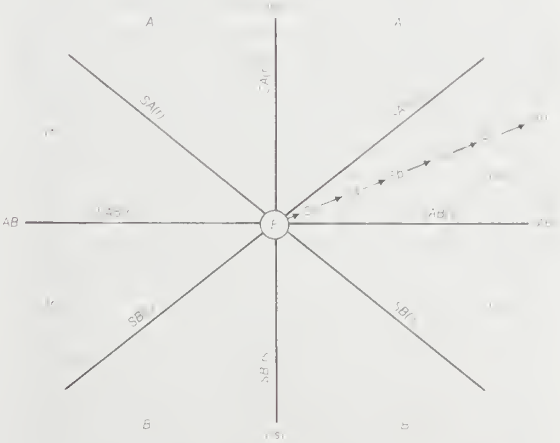


Figure 3: Gerard de Vaucouleurs's pinwheel diagram, a flat projection of his three-dimensional classification system.

Irregular galaxies. Hubble recognized two types of irregular galaxies, Irr I and Irr II. The Irr I type is the most common of the irregular systems, and it seems to fall naturally on an extension of the spiral classes, beyond Sc, into galaxies with no discernible spiral structure. They are blue, highly resolved, and have little or no nucleus. The Irr II systems are rare objects. They include various kinds of chaotic galaxies for which there apparently are many different explanations. Table 2 compares various subgroups of this rather confusing assembly of objects.

Table 2: Types of Irr II Galaxies

	dusty	star-forming	tidally interacting
Examples	NGC 4433 NGC 4753 NGC 5363	NGC 1569 NGC 4691 NGC 5253	NGC 520 NGC 3448 NGC 5195
Spectra	F2–GO	A–F	A–F
$(B-V)$ colour	0.9–1.2	0.3–0.6	0.7–1.0
Radio luminosity	normal	normal	strong
Diffuse emission	normal	strong	sometimes strong
H II regions	absent	present	present
Hα filaments	absent	present	absent
Dust content	very large	large	normal

Some irregular galaxies, like spirals, are barred. They have a nearly central bar structure dominating an otherwise chaotic arrangement of material. The Large Magellanic Cloud is a well-known example. The Hubble system does not normally recognize this as a subtype, though the de Vaucouleurs classification scheme includes it and its related types as Im and IB (see Table 1).

The Milky Way Galaxy

Large spiral system

The Milky Way Galaxy, sometimes simply called the Galaxy, is a spiral system consisting of several billion stars, one of which is the Sun. It takes its name from the Milky Way, the irregular luminous band of stars and gas clouds that stretches across the sky. Although the Earth lies well within the Galaxy, astronomers do not have as clear an understanding of its nature as they do of some external star systems. Because a thick layer of interstellar dust obscures much of the Galaxy from scrutiny by optical telescopes, astronomers can only determine its large-scale structure with the aid of radio and infrared telescopes, which can detect the forms of radiation that penetrate the obscuring matter.

SIZE AND MASS

As noted earlier, the first reliable measurement of the size of the Milky Way Galaxy was made in 1917 by Harlow Shapley. He arrived at his size determination by establishing the spatial distribution of globular clusters. Instead of a relatively small system with the Sun near its centre, as had previously been thought, Shapley found that the Galaxy was immense, with the Sun nearer the edge than the centre (Figure 4). Assuming that the globular clusters outlined the Galaxy, he determined that it has a diameter of about 100,000 light-years and that the Sun lies about 30,000 light-years from the centre. His values have held up remarkably well over the years. Although dependent in part on the particular component being discussed, with neutral hydrogen somewhat more widely dispersed and dark (*i.e.,* nonobservable) matter perhaps filling an even larger volume than expected, the stellar disk of the Milky Way system is just about as large as Shapley's model predicted. The most distant stars and gas clouds of the system that have had their distance determined reliably lie roughly 72,000 light-years from the galactic centre, while the distance of the Sun from the centre has been found to be approximately 27,000 light-years.

The total mass of the Galaxy, which had seemed reasonably well established during the 1960s, has become a matter of considerable uncertainty. Measuring the mass out to the distance of the farthest large hydrogen clouds is a relatively straightforward procedure. The measurements required are the velocities and positions of neutral hydrogen gas, combined with the approximation that the gas is rotating in nearly circular orbits around the centre of the Galaxy. A rotation curve, which relates the circular velocity of the gas to its distance from the galactic centre, is constructed. The shape of this curve and its values are determined by the amount of gravitational pull that the Galaxy exerts on the gas. Velocities are low in the central parts of the system because not much mass is interior to the orbit of the gas; most of the Galaxy is exterior to it and does not exert an inward gravitational pull. Velocities are high at intermediate distances because most of the mass in that case is inside the orbit of the gas clouds and the gravitational pull inward is at a maximum. At the farthest distances, the velocities decrease because nearly all the mass is interior to the clouds. This portion of the Galaxy is said to have Keplerian orbits, since the material should move in the same manner that the German astronomer Johannes Kepler discovered the planets to move within the solar system, where virtually all the mass is concentrated inside the orbiting bodies. The total mass of the Galaxy is then found by constructing mathematical models of the system with different amounts of material distributed in various ways and by comparing the resulting velocity curves with the observed one. As applied in the 1960s, this procedure indicated that the total mass of the Galaxy was approximately 200,000,000,000 times the mass of the Sun.

From A Blaauw and M Schmidt (eds), *Galactic Structure,* © by the University of Chicago, all rights reserved

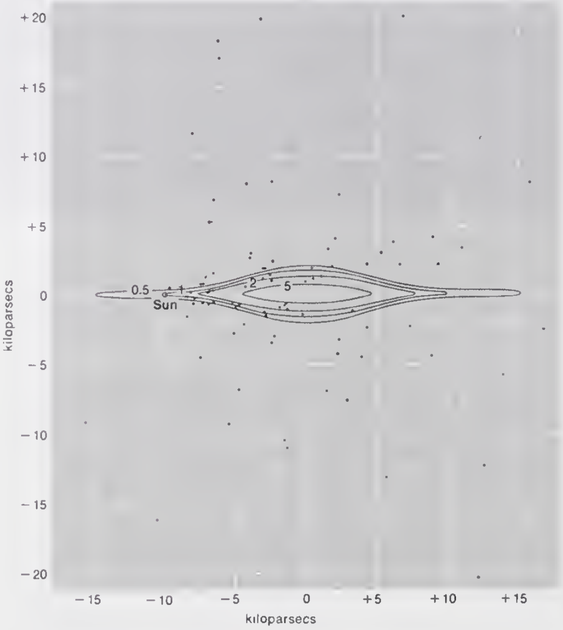


Figure 4: The distribution of globular clusters projected on a plane perpendicular to the galactic plane and passing through the galactic centre and the Sun. Lines of equal mass density are also drawn, but the high stellar density in the galactic centre is not indicated. North is at the top. Distances from the centre of the Galaxy in kiloparsecs are given below and to left of figure.

During the 1980s, however, refinements in the determination of the velocity curve began to cast doubts on the earlier results. The downward trend to lower velocities in the outer parts of the Galaxy was found to have been in error. Instead, the curve remained almost constant, indicating that there continue to be substantial amounts of matter exterior to the measured hydrogen gas. This in turn indicates that there must be some undetected material out there that is completely unexpected. It must extend considerably beyond the previously accepted positions of the edge of the Galaxy, and it must be dark at virtually all wavelengths, as it remains undetected even when searched for with radio, X-ray, ultraviolet, infrared, and optical telescopes. Until the dark matter is identified and its distribution determined, it will be impossible to measure the total mass of the Galaxy from the rotation curve, and so all that can be said is that the mass is

Dark matter

perhaps five or 10 times larger than thought earlier. That is to say, the mass, including the dark matter, must be about 1,000,000,000,000 times the mass of the Sun, with considerable uncertainty.

The nature of the dark matter in the Galaxy remains one of the major questions of galactic astronomy. Other galaxies also appear to have such matter in their outer parts. The only possible kinds of material that are consistent with the nondetections are all rather unlikely, at least according to present understanding in physics and astronomy. Planets and rocks would be impossible to detect, but it is extremely difficult to understand how they could materialize in sufficient numbers in the outer parts of galaxies where there are no stars or even interstellar gas and dust from which they could be formed. Massive neutrinos and other exotic, hypothetical subatomic particles also might be difficult to detect, but there is no good evidence that they even exist, and therefore they can only be considered a highly conjectural solution to the puzzle. It will take considerable effort to identify the dark matter with any degree of certainty. In the meantime it must be said that astronomy does not know what makes up much of the universe.

MAJOR COMPONENTS

Stars and stellar populations. The concept of different populations of stars has undergone considerable change over the last several decades. Before the 1940s astronomers were aware of differences among stars and had largely accounted for most of them in terms of different masses, luminosities, and orbital characteristics around the Galaxy. Understanding of evolutionary differences, however, had not yet been achieved, and differences in the chemical abundances in the stars were known but their significance was not comprehended. At this juncture chemical differences seemed exceptional and erratic and remained uncorrelated with other stellar properties. There was still no systematic division of stars even into different kinematic families in spite of the advances in theoretical work on the dynamics of the Galaxy.

Principal population types. In 1944 Baade announced the successful resolution into stars of the centre of the Andromeda Galaxy, M31, and its two elliptical companions, M32 and NGC 205. He found that the central parts of Andromeda and the accompanying galaxies were resolved at very much fainter magnitudes than were the outer spiral arm areas of M31. Furthermore, by using plates of different spectral sensitivity and coloured filters, he discovered that the two ellipticals and the centre of the spiral had red giants as their brightest stars rather than blue main-sequence stars, as in the case of the spiral arms. This finding led Baade to suggest that these galaxies, and also the Milky Way Galaxy, are made of two populations of stars that are distinct in their physical properties as well as their locations. He applied the term Population I to the stars that constitute the spiral arms of Andromeda and to most of the stars that are visible in the Milky Way system in the neighbourhood of the Sun. He found that these Population I objects were limited to the flat disk of the spirals and suggested that they were absent from the centres of such galaxies and from the ellipticals entirely. Baade designated as Population II the bright red giant stars that he discovered in the ellipticals and in the nucleus of Andromeda. Other objects that seemed to contain the brightest stars of this class were the globular clusters of the Galaxy. Baade further suggested that the high-velocity stars near the Sun (see below) were Population II objects that happened to be passing through the disk.

Population I and Population II

As a result of Baade's pioneering work on other galaxies in the Local Group (the cluster of star systems to which the Milky Way Galaxy belongs), astronomers immediately applied the notion of two stellar populations to the Galaxy. It is possible to segregate various components of the Galaxy into the two population types by applying both the idea of kinematics of different populations suggested by their position in the Andromeda system and the dynamical theories that relate galactic orbital properties with z distances (the distances above the plane of the Galaxy) for different stars. For many of these objects, the kinematic data on velocities are the prime source of population classification. The Population I component of the Galaxy, highly limited to the flat plane of the system, contains such objects as open star clusters, O and B stars, Cepheid variables, emission nebulas, and neutral hydrogen. Its Population II component, spread over a more nearly spherical volume of space, includes globular clusters, RR Lyrae variables, high-velocity stars, and certain other rarer objects (Figure 5).

From A Blaauw and M Schmidt (eds), *Galactic Structure*. © by the University of Chicago, all rights reserved

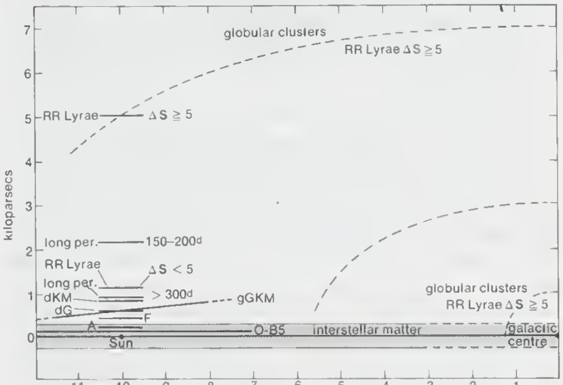


Figure 5: The space distribution of representative objects for various star populations projected on a plane perpendicular to the galactic plane and passing through the centre and the Sun. The horizontal lines indicate the distance from the galactic plane at which the density has fallen to one-tenth of that in the plane. *R* is the distance from the galactic centre. The RR Lyrae stars with high "ΔS" values have lowest metal content. The dashed lines indicate schematic (tentative) distributions (see text).

As time progressed, it was possible for astronomers to subdivide the different populations in the Galaxy further. Table 3 summarizes the properties and membership of the five subdivisions that were accepted at the time of the Vatican Conference on stellar populations in 1957. These subdivisions ranged from the nearly spherical "halo Population II" system to the very thin "extreme Population I" system, and each of the subgroups was found to contain (though not exclusively) characteristic types of stars. It was even possible to divide some of the variable-star types into subgroups according to their population subtype. The RR Lyrae variables of type ab, for example, could be separated into different groups by their spectral classifications and their mean periods. Those with mean periods longer than 0.4 days were classified as "halo Population II," while those with periods less than 0.4 days were placed in the "disk population." Similarly, long-period variables were divided into different subgroups, such that those with periods of less than 250 days and of relatively early spectral type (earlier than M 5e) were considered "intermediate Population II," whereas the longer period variables fell into the "older Population I" category.

An understanding of the physical differences in the stellar populations became increasingly clearer during the 1950s with improved calculations of stellar evolution. Evolving-star models showed that giants and supergiants are evolved objects recently derived from the main sequence after the exhaustion of hydrogen in the stellar core. As this became better understood, it was found that the luminosity of such giants was not only a function of the masses of the initial main-sequence stars from which they evolved but was also dependent on the chemical composition of the stellar atmosphere. Therefore, not only was the existence of giants in the different stellar populations understood but differences among the giants with relation to the main sequence of star groups came to be understood in terms of the chemistry of the stars.

At the same time, progress was made in determining the abundances of stars of the different population types by means of high-dispersion spectra obtained with large reflecting telescopes having a coudé focus arrangement. A curve of growth analysis demonstrated beyond a doubt that the two population types exhibited very different chemistries. In 1959 H. Lawrence Helfer, George Wallerstein, and Jesse L. Greenstein of the United States showed that the giant stars in globular clusters have chemical abundances quite different from those of Population I stars such as typified by the Sun. Population II stars have considerably lower abundances of the heavy elements—by amounts ranging from a factor of five or 10 up to a factor of several hundred. The total abundance of heavy elements, Z, for typical Population I stars is 0.04 (given in terms of the mass percent for all elements with atomic weights heavier than helium, a common practice in calculating stellar models). The values of Z for halo population globular clusters, on the other hand, were typically as small as 0.003.

Age estimates of stellar populations

A further difference between the two populations became clear as the study of stellar evolution advanced. It was found that Population II was exclusively made up of stars that are very old. Estimates of the age of Population II stars have varied over the years, depending on the degree of sophistication of the calculated models and the manner in which observations for globular clusters are fitted to these models. They have ranged from 10^9 years up to 2×10^{10} years. Recent comparisons of these data suggest that the halo globular clusters have ages of approximately 1.6×10^{10} years. The work of Sandage and his collaborators proved without a doubt that the range in age for globular clusters was relatively small and that the detailed characteristics of the giant branches of their colour-magnitude diagrams were correlated with age and small differences in chemical abundances. On the other hand, stars of Population I were found to have a wide range of ages. Stellar associations and galactic clusters with bright blue main-sequence stars have ages of a few million years (stars are still in the process of forming in some of them) to 10^{10} years or more. Studies of the stars nearest the Sun indicate a mixture of ages with a considerable number of stars of great age—on the order of 10^9 years. Careful searches, however, have shown that there are no stars in the solar neighbourhood and no galactic clusters whatsoever that are older than or even quite as old as the globular clusters. This is an indication that globulars, and thus Population II objects, formed first in the Galaxy and that Population I stars have been forming since.

In short, as the understanding of stellar populations grew, the division into Population I and Population II became understood in terms of three parameters: age, chemical composition, and kinematics. A fourth parameter, spatial distribution, appeared to be clearly another manifestation of kinematics. The correlations between these three parameters were not perfect but seemed to be reasonably good for the Galaxy, even though it was not yet known whether these correlations were applicable to other galaxies. Table 3 illustrates the close correlations, as formulated in the early 1960s, for the stars in the Galaxy and shows that there are many different combinations of these three parameters that seem to be excluded in nature. The many different physical manifestations of these parameters were gradually building up. Methods of determining the abundance of metals in objects by means other than laborious high-dispersion coudé spectroscopy became possible. For example, it was found that stars having a low abundance of heavy elements exhibited an easily measurable ultraviolet excess. This is demonstrated when the three colours U, B, and V of the Yerkes system are plotted in a three-colour diagram where the Population II stars all lie distinctly to the left of the normal star, three-colour relationship.

Understanding the evolution of the Galaxy's stellar population

Astronomers devised a graphic way to explain the evolution of the stellar population in the Milky Way Galaxy using a three-dimensional plot in which the age, the abundance of heavy elements, and the rate of star formation are all taken into account. Figure 6 is an example of such a three-dimensional plot. The volume shown in the figure indicates that the rate of star formation about the time the Galaxy originated was somewhat greater than at present but that it has not yet reached zero. As stars formed, the heavy elements were produced in the hot centres of the stars and in supernovas; thus the volume moves forward in the box until the present is reached, and the majority of stars that are now forming have heavy elements in approximately the same amount as the Sun. At any time, τ, there is a spread in the abundances of the stars formed, depending on the history of the interstellar material in the region.

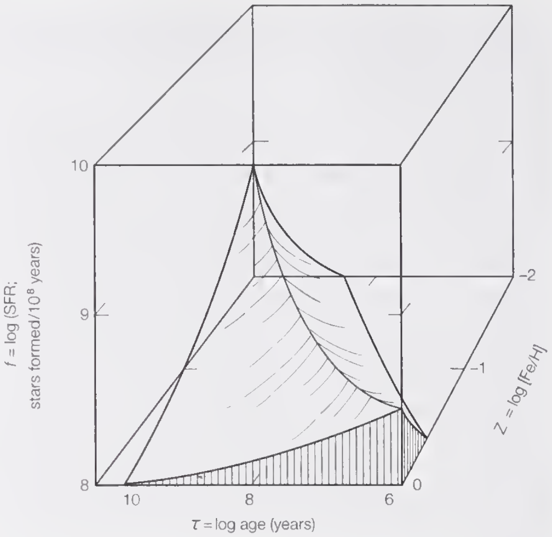


Figure 6: The star-formation history of the Milky Way Galaxy. The front axis is the age of the stars, the side axis is the heavy-element abundance of the stars (expressed in terms of the iron [Fe] abundance as compared with that of the Sun), and the vertical axis is the rate of stars that were being formed (SFR) at the corresponding time in the life of the Galaxy.

Complications in scientific understanding of stellar populations first became serious when detailed colour-magnitude diagrams were obtained for star clusters in the Magellanic Clouds during the late 1950s. Arp's work on the clusters of the Small Magellanic Cloud showed that the correlations between the properties of populations found in the Galaxy apparently broke down when other galaxies were examined carefully. Arp suggested that the star clusters of young age that he had observed in the SMC might be examples of young Population II stars—*i.e.,* young stars having a low abundance of heavy elements. No such stars were known in the Galaxy. Similarly, Arp found anomalous colour-magnitude diagrams for globular clusters in the SMC and proposed that perhaps this also was the result of abundance differences between the SMC and the Galaxy. At first it appeared that these conclusions were based on detailed comparisons with evolutionary models; however, because of the lack of such models at the time of Arp's observations, it seemed clear that the young star clusters of the SMC were anomalous in many details and that these peculiarities could not easily be accounted for other than by differences in chemical composition. In succeeding years, and as more star clusters in the Magellanic Clouds were measured, investigators were able to make detailed comparisons with models and to conclude that the chemical differences between the Galaxy and the Clouds must be rather small. Many of the star clusters have colour-magnitude diagrams that nearly conform to models calculated on the basis of solar-type abundances. It also is true, however, that many clusters, including those measured with high-detection-efficiency equipment such as the charge-coupled device (CCD), show real differences from colour-magnitude diagrams of galactic clusters, and these differences are still not completely understood. The Andromeda Galaxy has many globular clusters that can be observed with large instruments, and these also show a wider variety of properties than expected on the basis of the local sample. Surveys of the spectra and colours of the Andromeda globulars have demonstrated that there is a considerable spread in heavy-element abundance for these systems and that the close correlation between position and abundance found in the Galaxy fails to materialize in the case of Andromeda. Consequently, the segregation into stellar populations that works so well for the Galaxy is not necessarily a universal system. Moreover, it is possi-

Table 3: The Stellar Populations

	Population I		disk population	Population II	
	extreme Population I	older Population I		intermediate Population II	halo Population II
Members	gas young stars associated with the present spiral structure supergiants Cepheids T Tauri stars galactic clusters of Trumpler's class I	A-type stars strong-line stars Me dwarfs	stars of galactic nucleus planetary nebulas novas RR Lyrae stars with periods < 0.4 day weak-line stars	high-velocity stars with z-velocities > 30 km/sec long-period variables with periods < 250 days and spectral types earlier than M5e	subdwarfs globular clusters RR Lyrae stars with periods longer than 0.4 days
Average height over galactic plane (parsecs)	120	160	400	700	2,000
Average velocity perpendicular to galactic plane z (km/sec)	8	10	17	25	75
Axial ratio of spheroidal distribution	100	?	25?	5	2
Concentration toward centre	little	little	strong?	strong	strong
Distribution	extremely patchy, spiral arms	patchy, spiral arms	smooth?	smooth	smooth
Age (10^9 years)	0.1	0.1 to 1.5	1.5 to 5.0	6.0 to 5.0	6
Total mass (10^9 suns)	2	5	47 (combined disk and intermediate)		16

ble that most of the correlations are connected specifically to the detailed history and evolution of the Galaxy rather than to fundamental properties that stars in general would be expected to possess.

The stellar luminosity function. The stellar luminosity function is a description of the relative number of stars of different absolute luminosities. It is often used to describe the stellar content of various parts of the Galaxy or other groups of stars, but it most commonly refers to the absolute number of stars of different absolute magnitudes in the solar neighbourhood. In this form it is usually called the van Rhijn function after the Dutch astronomer Pieter J. van Rhijn. The van Rhijn function is a basic datum for the local portion of the Galaxy, but it is not necessarily representative for an area larger than the immediate solar neighbourhood. Investigators have found that elsewhere in the Galaxy, and in the external galaxies (as well as in star clusters), the form of the luminosity function differs in various respects from the van Rhijn function.

The luminosity function of the solar neighbourhood

The detailed determination of the luminosity function of the solar neighbourhood is an extremely complicated process. Difficulties arise because of (1) the incompleteness of existing surveys of stars of all luminosities in any sample of space and (2) the uncertainties in the basic data (distances and magnitudes). In determining the van Rhijn function, it is normally preferable to specify exactly what volume of space is being sampled and to state explicitly the way in which problems of incompleteness and data uncertainties are handled.

In general there are four different methods for determining the local luminosity function. Most commonly, trigonometric parallaxes are employed as the basic sample. Alternative but somewhat less certain methods include the use of spectroscopic parallaxes, which can involve much larger volumes of space. A third method entails the use of mean parallaxes of a star of a given proper motion and apparent magnitude; this yields a statistical sample of stars of approximately known and uniform distance. The fourth method involves examining the distribution of proper motions and tangential velocities (the speeds at which stellar objects move at right angles to the line of sight) of stars near the Sun.

Because the solar neighbourhood is a mixture of stars of various ages and different types, it is difficult to interpret the van Rhijn function in physical terms without recourse to other sources of information, such as the study of star clusters of various types, ages, and dynamical families. Globular clusters are the best samples to use for determining the luminosity function of old stars having a low abundance of heavy elements (Population II stars).

Globular-cluster luminosity functions show a conspicuous peak at absolute magnitude $M_V = 0.5$, and this is clearly due to the enrichment of stars at that magnitude from the horizontal branch of the cluster. The height of this peak in the data is related to the richness of the horizontal branch, which is in turn related to the age and chemical composition of the stars in the cluster. A comparison of the observed M3 luminosity function with the van Rhijn function shows a depletion of stars, relative to fainter stars, for absolute magnitudes brighter than roughly $M_V = 3.5$. This discrepancy is important in the discussion of the physical significance of the van Rhijn function and luminosity functions for clusters of different ages, and so will be dealt with more fully below.

Luminosity functions of open clusters

Many studies of the component stars of open clusters have shown that the luminosity functions of these objects vary widely. The two most conspicuous differences are the overabundance of stars of brighter absolute luminosities and the underabundance or absence of stars of faint absolute luminosities. The overabundance at the bright end is clearly related to the age of the cluster (as determined from the main-sequence turnoff point) in the sense that younger star clusters have more of the highly luminous stars. This is completely understandable in terms of the evolution of the clusters and can be accounted for in detail by calculations of the rate of evolution of stars of different absolute magnitudes and mass. For example, the luminosity function for the young clusters h and χ Persei, when compared to the van Rhijn function, clearly shows a large overabundance of bright stars due to the extremely young age of the cluster, which is on the order of 10^6 years. Calculations of stellar evolution indicate that in an additional 10^9 or 10^{10} years all of these stars will have evolved away and disappeared from the bright end of the luminosity function.

In 1955 the first detailed attempt to interpret the shape of the general van Rhijn luminosity function was made by the Austrian-born astronomer Edwin E. Salpeter, who pointed out that the change in slope of this function near $M_V = +3.5$ is most likely the result of the depletion of the stars brighter than this limit. Salpeter noted that this particular absolute luminosity is very close to the turnoff point of the main sequence for stars of an age equal to the oldest in the solar neighbourhood—approximately 10^{10} years. Thus, all stars of the luminosity function with fainter absolute magnitudes have not suffered depletion of their numbers because of stellar evolution as there has not been enough time for them to have evolved from the main sequence. On the other hand, the ranks of stars of brighter absolute luminosity have been variously depleted by evolution, and so the form of the luminosity function in this range is a composite curve contributed by stars of

Formation function

ages ranging from 0 to 10^{10} years. Salpeter hypothesized that there might exist a time-independent function, the so-called formation function, which would describe the general initial distribution of luminosities, taking into account all stars at the time of formation. Then, by assuming that the rate of star formation in the solar neighbourhood has been uniform since the beginning of this process and by using available calculations of the rate of evolution of stars of different masses and luminosities, he showed that it is possible to apply a correction to the van Rhijn function in order to obtain the form of the initial luminosity function. Comparisons of open clusters of various ages have shown that these clusters agree much more closely with the initial formation function than with the van Rhijn function; this is especially true for the very young clusters. Consequently, investigators believe that the formation function, as derived by Salpeter, is a reasonable representation of the distribution of star luminosities at the time of formation, even though they are not certain that the assumption of a uniform rate of formation of stars can be precisely true or that the rate is uniform throughout a galaxy.

It was stated above that open-cluster luminosity functions show two discrepancies when compared with the van Rhijn function. The first is due to the evolution of stars from the bright end of the luminosity function such that young clusters have too many stars of high luminosity, as compared to the solar neighbourhood. The second discrepancy is that very old clusters such as the globulars have too few high-luminosity stars, as compared to the van Rhijn function, and this is clearly the result of stellar evolution away from the main sequence. Stars do not, however, disappear completely from the luminosity function; most become white dwarfs and reappear at the faint end. In his early comparisons of formation functions with luminosity functions of galactic clusters, Sandage calculated the number of white dwarfs expected in various clusters; present searches for these objects in a few of the clusters (*e.g.,* the Hyades) have supported his conclusions.

Open clusters also disagree with the van Rhijn function at the faint end—*i.e.,* for absolute magnitudes fainter than approximately $M_V = +6$. In all likelihood this is mainly due to a depletion of another sort, the result of dynamical effects on the clusters that arise because of internal and external forces. Stars of low mass in such clusters escape from the system under certain common conditions. The formation functions for these clusters may be different from the Salpeter function and may exclude faint stars. A further effect is the result of the finite amount of time it takes for stars to condense; very young clusters have few faint stars partly because there has not been sufficient time for them to have reached their main-sequence luminosity.

Star clusters and stellar associations. Although most stars in the Galaxy exist either as single stars like the Sun or as double stars, there are many conspicuous groups and clusters of stars that contain tens to thousands of members. These objects can be subdivided into three types: globular clusters, open clusters, and stellar associations. They differ primarily in age and in the number of member stars.

Principal types

Globular clusters. The largest and most massive star clusters are the globular clusters, so called because of their roughly spherical appearance (Figure 7). The Galaxy contains approximately 130 globular clusters (the exact number is uncertain because of obscuration by dust in the Milky Way band, which probably prevents some 10 or so globulars from being seen). They are arranged in a nearly spherical halo around the Milky Way, with relatively few toward the galactic plane but a heavy concentration toward the centre. The radial distribution, when plotted as a function of distance from the galactic centre, fits a mathematical expression of a form identical to the one describing the star distribution in elliptical galaxies, though there is an anomalous peak in the distribution at distances of about 40,000 light-years from the centre.

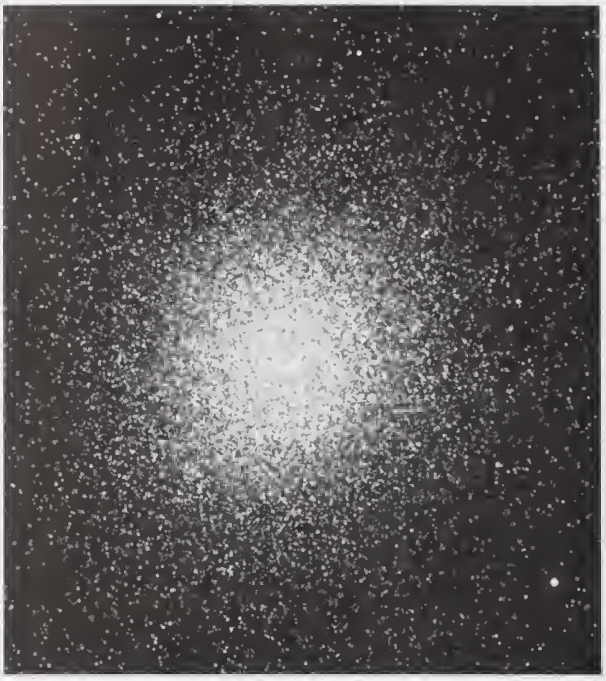


Figure 7: Omega Centauri (M3), the most imposing globular cluster in the Milky Way Galaxy.

By courtesy of the National Optical Astronomy Observatories

Globular clusters are extremely luminous objects. Their mean luminosity is the equivalent of approximately 25,000 suns. The most luminous are 50 times brighter. The masses of globular clusters, measured by determining the dispersion in the velocities of individual stars, range from a few thousand to more than 1,000,000 solar masses. The clusters are very large, with diameters measuring from 10 to as much as 300 light-years. Most globular clusters are highly concentrated at their centres, having stellar distributions that resemble isothermal gas spheres with a cutoff that corresponds to the tidal effects of the Galaxy. A precise model of star distribution within a cluster can be derived from stellar dynamics, which takes into account the kinds of orbits that stars have in the cluster, encounters between these member stars, and the effects of exterior influences. The American astronomer Ivan R. King, for instance, has derived dynamical models that fit observed stellar distributions very closely. He finds that a cluster's structure can be described in terms of two numbers: (1) the core radius, which measures the degree of concentration at the centre, and (2) the tidal radius, which measures the cutoff of star densities at the edge of the cluster.

A key distinguishing feature of globular clusters in the Galaxy is their uniformly old age. Determined by comparing the stellar population of globulars with stellar evolutionary models, the ages of all those so far measured range from 12 billion to 18 billion years. They are the oldest objects in the Galaxy and so must have been among the first formed when the system condensed out of the pregalactic gas. That this was the case is also indicated by the fact that the globulars tend to have much smaller amounts of heavy elements than do the stars in the plane of the Galaxy—*e.g.,* the Sun. Composed of stars belonging to the extreme Population II, as well as the high-latitude halo stars, these nearly spherical assemblages apparently formed before the material of the Galaxy flattened into the present thin disk. As their component stars evolved, they gave up some of their gas to interstellar space. This gas was enriched in the heavy elements produced in stars during the later stages of their evolution, so that the interstellar gas in the Galaxy is continually being changed. Hydrogen and helium have always been the major constituents, but heavy elements have gradually grown in importance. The present interstellar gas contains elements heavier than helium at a level of about 2 percent by mass, while the globulars contain as little as 0.02 percent of the same elements.

Open clusters. Clusters smaller and less massive than the globulars are found in the plane of the Galaxy intermixed with the majority of the system's stars, including the Sun. These objects are the open clusters, so called because they generally have a more open, loose appearance than typical globulars.

Open clusters are distributed in the Galaxy very similarly to young stars. They are highly concentrated along the plane of the Galaxy and slowly decrease in number outward from its centre. The large-scale distribution of these clusters cannot be learned directly because their existence in the Milky Way plane means that dust obscures those that are more than a few thousand light-years from the Sun. By analogy with open clusters in external galaxies similar to the Galaxy it is surmised that they follow the general distribution of integrated light in the Galaxy, except that there are probably fewer of them in the central areas. There is some evidence that the younger open clusters are more densely concentrated in the Galaxy's spiral arms, at least in the neighbourhood of the Sun where these arms can be discerned.

The brightest open clusters are considerably fainter than the brightest globular clusters. The peak absolute luminosity appears to be about 50,000 times the luminosity of the Sun, but the largest percentage of known open clusters has a brightness equivalent to 500 solar luminosities. Masses can be determined from the dispersion in the measured velocities of individual stellar members of clusters. Most open clusters have small masses—on the order of 50 solar masses. Their total populations of stars are small, ranging from tens to a few thousand.

Small mass of open clusters

Open clusters have diameters of only two or three to about 20 light-years, with the majority being less than five light-years across. In structure they look very different from globular clusters, though they can be understood in terms of similar dynamical models. The most important structural difference is their small total mass and relative looseness, which result from their comparatively large core radii. These two features have disastrous consequences as far as their ultimate fate is concerned, because open clusters are not sufficiently gravitationally bound to be able to withstand the disruptive tidal effects of the Galaxy (see STARS AND STAR CLUSTERS). Judging from the sample of open clusters within 3,000 light-years of the Sun, only half of them can withstand such tidal forces for more than 200,000,000 years, while a mere 2 percent have life expectancies as high as 1,000,000,000 years.

Measured ages of open clusters agree with the conclusions that have been reached about their life expectancies. They tend to be young objects; only a few are known to exceed 1,000,000,000 years in age. Most are younger than 200,000,000 years, and some are 1,000,000 or 2,000,000 years old. Ages of open clusters are determined by comparing their stellar membership with theoretical models of stellar evolution. Because all the stars in a cluster have very nearly the same age and chemical composition, the differences between the member stars are entirely the result of their different masses. As time progresses after the formation of a cluster, the massive stars, which evolve the fastest, gradually disappear from the cluster, becoming white dwarf stars or other underluminous stellar remnants. Theoretical models of clusters show how this effect changes the stellar content with time, and direct comparisons with real clusters give reliable ages for them. Astronomers use a diagram (the colour-magnitude diagram) that plots the temperatures of the stars against their luminosities to make this comparison. Colour-magnitude diagrams have been obtained for about 1,000 open clusters, and ages are thus known for this large sample.

Because open clusters are mostly young objects, they have chemical compositions that correspond to the enriched environment from which they formed. Most of them are like the Sun in their abundance of the heavy elements, and some are even richer. For instance, the Hyades, which compose one of the nearest clusters, have almost twice the abundance of heavy elements as the Sun.

Stellar associations. Even younger than open clusters, stellar associations are very loose groupings of stars that share a common place and time of origin but that are not generally tied closely enough together gravitationally to form a stable cluster. Stellar associations are limited strictly to the plane of the Galaxy and appear only in regions of the system where star formation is occurring, notably in the spiral arms. They are very luminous objects. The brightest are even brighter than the brightest globular clusters, but this is not because they contain more stars; instead it is the result of the fact that their constituent stars are very much brighter than the stars constituting globulars. The most luminous stars in stellar associations are very young stars of spectral types O and B. They have absolute luminosities as bright as any star in the Galaxy—on the order of 1,000,000 times the luminosity of the Sun. Such stars have very short lifetimes, only lasting a few million years. With luminous stars of this type there need not be very many to make up a highly luminous and conspicuous grouping. The total masses of stellar associations amount to only a few hundred solar masses, with the population of stars being in the hundreds or, in a few cases, thousands.

High luminosity of associations

The sizes of stellar associations are large; the average diameter of those in the Galaxy is about 700 light-years. Many are smaller, especially near the Sun, where they measure about 200 light-years across. In any case, stellar associations are so large and loosely structured that their self-gravitation is insufficient to hold them together, and in a matter of a few million years the members disperse into surrounding space, becoming separate and unconnected stars in the galactic field.

Moving groups. These objects are remote organizations of stars that share common measurable motions but do not form a noticeable cluster. This definition allows the term to be applied to a range of objects from the nearest gravitationally bound clusters to groups of widely spread stars with no apparent gravitational identity, which are discovered only by searching the catalogs for stars of common motion. Among the best known of the moving groups is the Hyades in the constellation Taurus. Also known as the Taurus moving cluster or the Taurus stream, this system is comprised of the relatively dense Hyades cluster, along with a few very distant members. It contains a total of about 350 stars, including several white dwarfs. Its centre lies about 150 light-years away. Other notable moving stellar groups include the Ursa Major, Scorpio-Centaurus, and Pleiades groups. Besides these remote organizations, investigators have observed what appears to be groups of high-velocity stars near the Sun. One of these, called the Groombridge 1830 group, consists of a number of subdwarfs and the star RR Lyrae after which the RR Lyrae variables were named.

Hyades moving cluster

Recent advances in the study of moving groups have had an impact on the investigation of the kinematic history of stars and on the absolute calibration of the distance scale of the Galaxy. Moving groups have proved particularly useful with respect to the latter because their commonality of motion enables astronomers to determine accurately (for the nearer examples) the distance of each individual member. Together with nearby parallax stars, moving-group parallaxes provide the basis for the galactic distance scale. Astronomers have found the Hyades moving cluster well suited for their purpose: it is close enough to permit the reliable application of the method, and it has enough members for deducing an accurate main-sequence position.

Moving-group parallaxes

One of the basic problems of using moving groups for distance determination is the selection of members. In the case of the Hyades this has been done very carefully but not without considerable dispute. The members of a moving group (and its actual existence) are established by the degree to which their motions define a common convergent point in the sky. One technique is to determine the coordinates of the poles of the great circles defined by the proper motions and positions of individual stars. The positions of the poles will define a great circle, and one of its poles will be the convergent point for the moving group. Membership of stars can be established by criteria applied to the distances of proper-motion poles of individual stars from the mean great circle. The reliability of the existence of the group itself can be measured by the dispersion of the great circle points about their mean.

As radial velocities will not have been used for the preliminary selection of members, they can be subsequently examined to eliminate further nonmembers. The final list of members should contain only a very few nonmembers—either those that appear to agree with the group motion because of observational errors or those that happen to share the group's motion at the present time but are not related to the group historically.

The distances of individual stars in a moving group may be determined if their radial velocities and proper motions are known (see below *Stellar motions*) and if the exact position of the radiant is determined. If the angular distance of a star from the radiant is λ, and if the velocity of the cluster as a whole with respect to the Sun is V, then the radial velocity of the star, V_r, is

$$V_r = V \cos \lambda.$$

The transverse (or tangential) velocity, T, is given by

$$T = V \sin \lambda = 4.74\mu/p,$$

where p is the star's parallax in arc seconds. Thus, the parallax of a star is given by

$$p = 4.74\mu \cot \lambda/V_r.$$

The key to achieving reliable distances by this method is to locate the convergent point of the group as accurately as possible. The various techniques used (*e.g.*, Charlier's method) are capable of high accuracy providing that the measurements themselves are free of systematic errors. For the Taurus moving group, for example, it has been estimated that the accuracy for the best observed stars is on the order of 3 percent in the parallax, discounting any errors due to systematic problems in the proper motions. By comparison, the trigonometric parallaxes of the same stars have errors of about 30 percent.

Emission nebulas. A conspicuous component of the Galaxy is the collection of large, bright, diffuse gaseous objects generally called nebulas. The brightest of these cloudlike objects are the emission nebulas, large complexes of interstellar gas and stars in which the gas exists in an ionized and excited state (with the electrons of the atoms excited to a higher than normal energy level). This condition is produced by the strong ultraviolet light emitted from the very luminous, hot stars embedded in the gas. Because emission nebulas consist almost entirely of ionized hydrogen, they are usually referred to as H II regions.

H II regions

H II regions are found in the plane of the Galaxy intermixed with young stars, stellar associations, and the youngest of the open clusters. They are areas where very massive stars have recently formed, and many contain the uncondensed gas, dust, and molecular complexes commonly associated with ongoing star formation. The H II regions are concentrated in the spiral arms of the Galaxy, though some do exist between the arms. Many of them are found at intermediate distances from the centre of the Milky Way, with the largest number occurring at a distance of 10,000 light-years. This latter fact can be ascertained even though the H II regions cannot be seen clearly beyond a few thousand light-years from the Sun. They emit radio radiation of a characteristic type, with a thermal spectrum that indicates that their temperatures are about 10,000 kelvins (K). This thermal radio radiation enables astronomers to map the distribution of H II regions in distant parts of the Galaxy.

The largest and brightest H II regions in the Galaxy rival the brightest star clusters in total luminosity. Even though most of the visible radiation is concentrated in a few discrete emission lines, the total apparent brightness of the brightest is the equivalent of tens of thousands of solar luminosities. These H II regions are also remarkable in size, having diameters of about 1,000 light-years. More typically, common H II regions such as the Orion Nebula are about 50 light-years across. They contain gas that has a total mass ranging from one or two solar masses up to several thousand. H II regions consist primarily of hydrogen, but they also contain measurable amounts of other gases. Helium is second in abundance, and large amounts of carbon, nitrogen, and oxygen occur as well. Preliminary evidence indicates that the ratio of the abundance of the heavier elements among the detected gases to hydrogen decreases outward from the centre of the Galaxy, a tendency that has been observed in other spiral galaxies.

Planetary nebulas. The gaseous clouds known as planetary nebulas are only superficially similar to other types of nebulas. So called because the smaller varieties almost resemble planetary disks when viewed through a telescope, planetary nebulas represent a stage at the end of the stellar life cycle rather than one at the beginning. The distribution of such nebulas in the Galaxy is different from that of H II regions. Planetary nebulas belong to an intermediate population and are found throughout the disk and the inner halo. There are slightly more than 1,000 known planetary nebulas in the Galaxy, but many might be overlooked because of obscuration in the Milky Way region.

Supernova remnants. Another type of nebulous object found in the Galaxy is the remnant of the gas blown out from an exploding star that forms a supernova. Occasionally these objects look something like planetary nebulas, as in the case of the Crab Nebula, but they differ from the latter in three ways: (1) the total mass of their gas (they involve a larger mass, essentially all the mass of the exploding star), (2) their kinematics (they are expanding with higher velocities), and (3) their lifetimes (they last for a shorter time as visible nebulas). The best-known supernova remnants are those resulting from three historically observed supernovas: that of AD 1054, which made the Crab Nebula its remnant; that of AD 1572, called Tycho's Nova; and that of AD 1604, called Kepler's Nova. These objects and the many others like them in the Galaxy are detected at radio wavelengths. They release radio energy in a nearly flat spectrum due to the emission of radiation by charged particles moving spirally at nearly the speed of light in a magnetic field enmeshed in the gaseous remnant. Radiation generated in this way is called synchrotron radiation and is associated with various types of violent cosmic phenomena besides supernova remnants, as, for example, radio galaxies.

Synchrotron radiation

Dust clouds. The dust clouds of the Galaxy are narrowly limited to the plane of the Milky Way, though very low-density dust can be detected even near the galactic poles. Dust clouds beyond 2,000 to 3,000 light-years from the Sun cannot be detected optically, because intervening clouds of dust and the general dust layer obscure more distant views. Based on the distribution of dust clouds in other galaxies, it can be concluded that they are often most conspicuous within the spiral arms, especially along the inner edge of well-defined ones. The best observed dust clouds near the Sun have masses of several hundred solar masses and sizes ranging from a maximum of about 200 light-years to a fraction of a light-year. The smallest tend to be the densest, possibly due in part to evolution: as a dust complex contracts, it also becomes denser and more opaque. The very smallest dust clouds are the so-called Bok globules, named after the Dutch-American astronomer Bart J. Bok; these objects are about one light-year across and have masses of one to 20 solar masses.

More complete information on the dust in the Galaxy comes from infrared observations. While optical instruments can detect the dust when it obscures more distant objects or when it is illuminated by very nearby stars, infrared telescopes are able to register the long-wavelength radiation that the cool dust clouds themselves emit. A complete survey of the sky at infrared wavelengths made during the early 1980s by an unmanned orbiting observatory, the Infrared Astronomy Satellite (IRAS), revealed a large number of dense dust clouds in the Milky Way.

Thick clouds of dust in the Milky Way can be studied by still another means. Many such objects contain detectable amounts of molecules that emit radio radiation at wavelengths that allow them to be identified and analyzed. More than 50 different molecules, including carbon monoxide and formaldehyde, and radicals have been detected in dust clouds.

The general interstellar medium. The stars in the Galaxy, especially along the Milky Way, reveal the presence of a general, all-pervasive interstellar medium by the way in which they gradually fade with distance. This occurs primarily because of interstellar dust, which obscures and reddens starlight. On the average, stars near the Sun are dimmed by a factor of two for every 3,000 light-years. Thus, a star that is 6,000 light-years away in the plane of the Galaxy will appear four times fainter than it would otherwise were it not for the interstellar dust.

Another way in which the effects of interstellar dust become apparent is through the polarization of background starlight. Dust is aligned in space to some extent, and this results in selective absorption such that there is a preferred plane of vibration for the light waves. The electric vectors tend to lie preferentially along the galactic plane, though there are areas where the distribution is more complicated. It is likely that the polarization arises because the dust grains are partially aligned by the galactic magnetic field. If the dust grains are paramagnetic so that they act somewhat like a magnet, then the general magnetic field, though very weak, can in time line up the grains with their short axes in the direction of the field. As a consequence, the directions of polarization for stars in different parts of the sky make it possible to plot the direction of the magnetic field in the Milky Way.

Interstellar neutral hydrogen

The dust is accompanied by gas, which is thinly dispersed among the stars, filling the space between them. This interstellar gas consists mostly of hydrogen in its neutral form. Radio telescopes can detect neutral hydrogen because it emits radiation at a wavelength of 21 centimetres. Such radio wavelength is long enough to penetrate interstellar dust and so can be detected from all parts of the Galaxy. Most of what astronomers have learned about the large-scale structure and motions of the Galaxy has been derived from the radio waves of interstellar neutral hydrogen. The distance to the gas detected is not easily determined. Statistical arguments must be used in many cases, but the velocities of the gas, when compared with the velocities found for stars and those anticipated on the basis of the dynamics of the Galaxy, provide useful clues as to the location of the different sources of hydrogen radio emission. Near the Sun the average density of interstellar gas is 10^{-21} gm/cm^3, which is the equivalent of about one hydrogen atom per cubic centimetre.

Even before they first detected the emission from neutral hydrogen in 1951, astronomers were aware of interstellar gas. Minor components of the gas, such as sodium and calcium, absorb light at specific wavelengths (see Table 4), and they thus cause the appearance of absorption lines in the spectra of the stars that lie beyond the gas. Since the lines originating from stars are usually different, it is possible to distinguish the lines of the interstellar gas and to measure both the density and velocity of the gas. Frequently it is even possible to observe the effects of several concentrations of interstellar gas between the Earth and the background stars and thereby determine the kinematics of the gas in different parts of the Galaxy.

STRUCTURE AND DYNAMICS

The structure of the Galaxy. The Galaxy's structure is fairly typical of a large spiral system. It can be viewed as consisting of six separate parts: (1) nucleus, (2) central bulge, (3) disk, (4) spiral arms, (5) spherical component, and (6) massive halo. Some of these components blend into each other; their differences in stellar population have been discussed above.

Sagittarius A

The nucleus. At the very centre of the Galaxy lies a remarkable object—in all likelihood a massive black hole surrounded by an accretion disk of high-temperature gas. Neither the central object nor any of the material immediately around it can be observed at optical wavelengths because of the thick screen of intervening dust in the Milky Way. The object, however, is readily detectable at radio wavelengths and has been dubbed Sagittarius A by radio astronomers. Somewhat similar to the centres of active galaxies (see below), though on a lesser scale, the galactic nucleus is the site of a wide range of activity apparently powered by the black hole. Infrared radiation and X rays are emitted from the area, and rapidly moving gas clouds can be observed there. Data strongly indicate that material is being pulled into the black hole from outside the nuclear region, including some gas from the z direction (*i.e.,* perpendicular to the galactic plane). As the gas nears the black hole, its strong gravitational force squeezes the gas into a rapidly rotating disk, which extends outward about five to 30 light-years from the central object. Rotation measurements of the disk indicate that the black hole has a mass roughly 4,000,000 times that of the Sun.

The central bulge. Surrounding the nucleus is an extended bulge of stars that is nearly spherical in shape and that consists primarily of Population II stars, though they are comparatively rich in heavy elements. Mixed with the stars are several globular clusters of similar stars, and both the stars and clusters have nearly radial orbits around the nucleus. The bulge stars can be seen optically where they stick up above the obscuring dust of the galactic plane.

Table 4: Interstellar Absorption Features

atomic lines		molecular lines		diffuse lines
atom or ion	λ (Å)	molecule	λ (Å)	λ (Å)
Na I	3302.34	CH	4300.31	4430.6
	3302.94		3890.23	4760
	5889.95		3886.39	5780.5
	5895.92		3878.77	5797.1
			3146.01	6203.0
K I	7664.91		3143.15	6270.0
	7698.98		3137.53	6283.9
				6613.9
Ca I	4226.73	CN	3874.61	
			3875.77	
Ca II	3933.66		3874.00	
	3968.47			
		CH+	4232.58	
Ti II	3072.97		3957.74	
	3229.19		3745.33	
	3241.98			
	3283.76			
Fe I	3719.94			
	3859.91			

The disk. From a distance the most conspicuous part of the Galaxy would be the disk, which extends from the nucleus out to distances of approximately 75,000 light-years. The Galaxy resembles other spiral systems, featuring as it does a bright, flat arrangement of stars and gas clouds that is spread out over its entirety and marked by a spiral structure. The disk can be thought of as being the underlying body of stars upon which the arms are superimposed. This body has a thickness that is roughly one-fifth its diameter, but different components have different characteristic thicknesses, as described below.

The spiral arms. Astronomers did not know that the Galaxy had a spiral structure until 1953, when the distances to stellar associations were first obtained reliably. Because of the obscuring interstellar dust and the interior location of the solar system, the spiral structure is very difficult to detect optically. This structure is easier to discern from radio maps of either neutral hydrogen or molecular clouds since both can be detected through the dust. Distances to the observed neutral hydrogen atoms must be estimated on the basis of measured velocities used in conjunction with a rotation curve for the Galaxy, which can be built up from measurements made at different galactic longitudes (see Figure 8).

From W. Becker and G. Contopoulos (eds.) *The Spiral Structure of the Galaxy* (1970) © Reidel Publishing Co. Dordrecht Holland

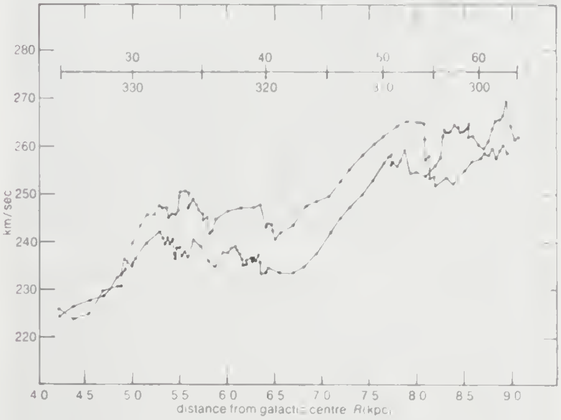


Figure 8: Galactic rotation curves from the 21-cm hydrogen emission line observations. Top curve refers to the gas north of (above) the galactic plane, bottom curve to points south of (below) the plane. The scale is set by the Sun's distance from the centre of the Galaxy, here taken as 10 kiloparsecs. The quantities *l* indicated above the curves are galactic longitudes.

From studies of other galaxies it can be shown that spiral arms generally follow a logarithmic spiral form such that

$$\log r = a - b\varphi,$$

where φ is a position angle measured from the centre to the outermost part of the arm, r is the distance from the centre of the galaxy, and a and b are constants. The range in pitch angles for galaxies is from about 50° to approximately 85°. The pitch angle is constant for any given galaxy if it follows a true logarithmic spiral. The pitch angle for the spiral arms of the Galaxy is difficult to determine from the limited optical data, but most measurements indicate a value of about 75°. There are five optically identified spiral arms in the part of the Milky Way wherein the solar system is located (see Table 5).

Table 5: Galactic Spiral Arms

arm number	name	galactic longitudes	distance
+I	Perseus	103–190°	10,000 light-years at 120°
0	Orion	60–250°	1,500 light-years at 180°
−I	Sagittarius	275–30°	5,000 light-years at 330°
−II	Norma	300–330°	10,000 light-years at 330°
−III	3kpc Arm	330–30°	20,000 light-years at 0°

Density-wave theory

Theoretical understanding of the Galaxy's spiral arms has progressed greatly in the 1980s, but there is still no complete understanding of the relative importance of the various effects thought to determine their structure. The overall pattern is almost certainly the result of a general dynamical effect known as a density-wave pattern. The Chinese-American astronomers Chia Chiao Lin and Frank H. Shu showed that a spiral shape is a natural result of any large-scale disturbance of the density distribution of stars in a galactic disk. When the interaction of the stars with one another is calculated, it is found that the resulting density distribution takes on a spiral pattern that does not rotate with the stars but rather moves around the nucleus more slowly as a fixed pattern. Individual stars in their orbits pass in and out of the spiral arms, slowing down in the arms temporarily and thereby causing the density enhancement. For the Galaxy, comparison of neutral hydrogen data with the calculations of Lin and Shu have shown that the pattern speed is 4 km/sec per 1,000 light-years.

Other effects that can influence a galaxy's spiral shape have been explored. It has been demonstrated, for example, that a general spiral pattern will result simply from the fact that the galaxy has differential rotation—*i.e.,* the rotation speed is different at different distances from the galactic centre. Any disturbance, such as a sequence of stellar formation events that are sometimes found drawn out in a near-linear pattern, will eventually take on a spiral shape simply because of the differential rotation. Distortions that also can be included are the results of massive explosions such as supernova events. These, however, tend to have only fairly local effects.

The spherical component. The space above and below the disk of the Galaxy is occupied by a thinly populated extension of the central bulge. Nearly spherical in shape, this region is populated by the outer globular clusters, but it also contains many individual field stars of extreme Population II, such as RR Lyrae variables and dwarf stars deficient in the heavy elements. Structurally, the spherical component resembles an elliptical galaxy, following the same simple mathematical law of density with distance from the centre.

The massive halo. The most controversial and least understood component of the Galaxy is the presumed giant massive halo that is exterior to the entire visible part. As explained above, the existence of the massive halo is demonstrated by its effect on the outer rotation curve of the Galaxy. All that can be said with any certainty is that the halo extends considerably beyond a distance of 100,000 light-years from the centre and that its mass is five or 10 times greater than the mass of the rest of the Galaxy taken together. It is not known what its shape is, what its constituents are, or how far into intergalactic space it extends.

Density distribution. *The stellar density near the Sun.* The density distribution of stars near the Sun can be used to calculate the mass density of material (in the form of stars) at the Sun's distance within the Galaxy. It is therefore of interest not only from the point of view of stellar statistics but also in relation to galactic dynamics. In principle, the density distribution can be calculated by integrating the stellar luminosity function. In practice, because of uncertainties in the luminosity function at the faint end and because of variations at the bright end, the local density distribution is not simply derived nor is there agreement between different studies in the final result.

Stellar density determinations

In the vicinity of the Sun, stellar density can be determined from the various surveys of nearby stars and from estimates of their completeness. For example, Peter van de Kamp's investigation of stars within 17 light-years can be used to determine the density of stars in this volume of space. Similarly, Wilhelm Gliese's catalog of stars closer than 65 light-years can be used for a larger volume of space and a larger sampling of stars.

The result of van de Kamp's determination of star density shows that even in his extremely close sample, incompleteness remains a problem in the surveys. The 60 stars that he includes fill a volume of space equal to 20,000 cubic light-years. Therefore, it is possible to calculate that the stellar density is 0.003 stars per cubic light-year. This figure, however, includes double and multiple star systems in addition to single stars, and if only the number of the systems is used stellar density is reduced to 0.002 objects per cubic light-year.

From Gliese's catalog, which contains 1,049 stars in a volume with a radius of 65 light-years, the average calculated density is less than one-third the calculated density for stars within 17 light-years. Thus, this catalog is incomplete, and its incompleteness is probably attributable to the fact that it is more difficult to detect the faintest stars and faint companions in the more distant systems.

In short, the true density of stars in the solar neighbourhood is difficult to establish. The value most commonly quoted is 0.003 stars per cubic light-year, a value obtained by integrating the van Rhijn luminosity function with a cutoff taken $M = 14.3$. This is, however, distinctly smaller than the true density as calculated for the most complete sampling volume discussed above and is therefore an underestimate. Gliese has estimated that when incompleteness of the catalogs is taken into account, the true stellar density is on the order of 0.004 stars per cubic light-year, which includes the probable number of unseen companions of multiple systems.

The density distribution of stars can be combined with the luminosity–mass relationship to obtain the mass density in the solar neighbourhood, which includes only stars and not interstellar material. This mass density is 4×10^{-24}g/cm^3.

Density distribution of various types of stars. To examine what kinds of stars contribute to the overall density distribution in the solar neighbourhood, various statistical sampling arguments can be applied to catalogs and lists of stars. The result of such a procedure is summarized in Table 6, which lists some of the kinds of objects and gives the calculated mean density over an appropriate volume centred on the Sun. For rare objects such as globular clusters, the volume of the sample must of course be rather large compared to that required to calculate the density for more common stars. Note that the figures are given in terms of mass density rather than number density. Number density for clusters obviously is very much smaller than mass densities.

The predominance of dwarf M stars

The most common stars and those that contribute the most to the local stellar mass density are the dwarf M stars, which provide a total of 0.0008 solar mass per cubic light-year. It is interesting to note that RR Lyrae variables and planetary nebulas—though many are known and thoroughly studied—contribute almost imperceptibly to the local star density. At the same time, white dwarfs, which are difficult to observe and of which very few are known, are among the more significant contributors.

Variations in the stellar density. The star density in the solar neighbourhood is not perfectly uniform. The most

Table 6: Space Densities of Stars

object	density (solar mass per cubic light-year)
O, B stars	0.00003
A, F stars	0.0001
dG, dK stars	0.0004
dM stars	0.0008
gG, gK stars	0.00003
gM stars	0.0000003
Dark companions	0.00014
White dwarfs	0.0002
Long-period variables	0.00000003
RR Lyrae stars	0.000000003
Cepheids	0.00000003
Planetary nebulas	0.0000000015
Open clusters	0.0000011
Globular clusters	0.00000003

conspicuous variations occur in the *z* direction, above and below the plane of the Galaxy, where the number density falls off rapidly. This will be considered separately below. The more difficult problem of variations within the plane is dealt with here.

Density variations are conspicuous for early type stars (*i.e.*, stars of higher temperatures) even after allowance has been made for interstellar absorption. For the stars earlier than type B3, for example, large stellar groupings in which the density is abnormally high are conspicuous in several galactic longitudes. The Sun in fact appears to be in a somewhat lower density region than the immediate surroundings where early B stars are relatively scarce. There is a conspicuous grouping of stars, sometimes called the Cassiopeia–Taurus Group, that has a centroid at approximately 600 light-years distance. A deficiency of early type stars is readily noticeable, for instance, in the direction of the constellation Perseus at distances beyond 600 light-years. Of course, the nearby stellar associations are striking density anomalies for early type stars in the solar neighbourhood. The early type stars within 2,000 light-years are significantly concentrated at negative galactic latitudes. This is a manifestation of a phenomenon referred to as "Gould's Belt," a tilt of the nearby bright stars in this direction with respect to the galactic plane. First noted by the English astronomer John Herschel in 1847, such anomalous behaviour is true only for the immediate neighbourhood of the Sun; faint B stars are strictly concentrated along the galactic equator.

Generally speaking, the large variations in stellar density near the Sun are less conspicuous for the late type dwarf stars (those of lower temperatures) than for the earlier types. This fact is explained as the result of the mixing of stellar orbits over long time intervals available for the older stars, which are primarily those stars of later spectral types. The young stars (O, B, and A types) are still close to the areas of star formation and show a common motion and common concentration due to initial formation distributions. In this connection it is interesting to note that the concentration of A-type stars at galactic longitudes 160° to 210° is coincident with a similar concentration of hydrogen detected by means of 21-centimetre line radiation. Correlations between densities of early type stars on the one hand and interstellar hydrogen on the other are conspicuous but not fixed; there are areas where neutral-hydrogen concentrations exist but for which no anomalous star density is found.

The variations discussed above are primarily small-scale fluctuations in star density rather than the large-scale phenomena so strikingly apparent in the structure of other galaxies. Sampling is too difficult and too limited to detect the spiral structure from the variations in the star densities for normal stars, although a hint of the spiral structure can be seen in the distribution in the earliest type stars and stellar associations. In order to determine the true extent in the star-density variations corresponding to these large-scale structural features, it is necessary to turn either to theoretical representations of the spiral structures or to other galaxies. From the former it is possible to find estimates of the ratio of star densities in the centre of spiral arms and in the interarm regions. The most commonly accepted theoretical representation of spiral structure, that of the density-wave theory, suggests that this ratio is on the order of 0.6, but for a complicated and distorted spiral structure such as apparently occurs in the Galaxy, there is no confidence that this figure corresponds very accurately with reality. On the other hand, fluctuations in other galaxies can be estimated from photometry of the spiral arms and the interarm regions provided that some indication of the nature of this stellar luminosity function at each position is available from colours or spectrophotometry. Estimates of the star density measured across the arms of spiral galaxies and into the interarm regions show that the large-scale spiral structure of a galaxy of this type is, at least in many cases, represented by only a relatively small fluctuation in star density.

Extremes in star density

It is clear from studies of the external galaxies that the range in star densities existing in nature is immense. For example, the density of stars at the centre of the nearby Andromeda spiral galaxy has been determined to equal 100,000 solar masses per cubic light-year, while the density at the centre of the Ursa Minor dwarf elliptical galaxy is only 0.00003 solar mass per cubic light-year.

Variation of star density with z distances. For all stars, variation of star density above and below the galactic plane rapidly decreases with height. Stars of different types, however, exhibit widely differing behaviour in this respect, and this tendency is one of the important clues as to the kinds of stars that occur in different stellar populations (see Table 3).

The luminosity function of stars is different at different galactic latitudes, and this is still another phenomenon connected with the *z* distribution of stars of different types. At a height of $z = 3{,}000$ light-years, stars of absolute magnitude 13 and fainter are nearly as abundant as at the galactic plane, while stars with absolute magnitude 0 are depleted by a factor of 100.

The values of the scale height for various kinds of objects given in Table 3 forms the basis for the segregation of these objects into different population types. Such objects as open clusters and Cepheid variables that have very small values of the scale height are the objects most restricted to the plane of the Galaxy, while globular clusters and other extreme Population II objects have scale heights of thousands of parsecs, indicating little or no concentration at the plane. Such data and the variation of star density with *z* distance bear on the mixture of stellar orbit types. They show the range from those stars having nearly circular orbits that are strictly limited to a very flat volume centred at the galactic plane to stars with highly elliptical orbits that are not restricted to the plane.

Basic elements of stellar motion

Stellar motions. A complete knowledge of a star's motion in space is possible only when both its proper motion and radial velocity can be measured. Proper motion is the motion of a star across an observer's line of sight and constitutes the rate at which the direction of the star changes in the celestial sphere. It is usually measured in seconds of arc per year. Radial velocity is the motion of a star along the line of sight and as such is the speed with which the star approaches or recedes from the observer. It is expressed in kilometres per second and given as either a positive or negative figure, depending on whether the star is moving away from or toward the observer.

Astronomers are able to measure both the proper motions and radial velocities of stars lying near the Sun. They can, however, determine only the radial velocities of stellar objects in more distant parts of the Galaxy and so must use these data, along with the information gleaned from the local sample of nearby stars, to ascertain the large-scale motions of stars in the Milky Way system.

Proper motions. The proper motions of the stars in the immediate neighbourhood of the Sun are usually very large, as compared to those of most other stars. Those of stars within 17 light-years of the Sun, for instance, range from 0.49 to 10.31 arc seconds per year. The latter value is that of Barnard's star, which is the star with the largest known proper motion. The tangential velocity of Barnard's star is 90 km/sec and, from its radial velocity (−108 km/sec) and distance (six light-years), astronomers have found that its space velocity (total velocity with respect to the Sun) is 140 km/sec. The distance to this star is rapidly decreasing; it will reach a minimum value of 3.5 light-years in about the year AD 11,800.

Radial velocities. Radial velocities, measured along the line of sight spectroscopically using the Doppler effect, are not known for all of the recognized stars near the Sun. Of the 45 systems within 17 light-years, only 40 have well-determined radial velocities. The radial velocities of the rest are not known either because of faintness or because of problems resulting from the nature of their spectrum. For example, radial velocities of white dwarfs are often very difficult to obtain because of the extremely broad and faint spectral lines in some of these objects. Moreover, the radial velocities that are determined for such stars are subject to further complication because a gravitational redshift generally affects the positions of their spectral lines. The average gravitational redshift for white dwarfs has been shown to be the equivalent of a velocity of −51 km/sec. To study the true motions of these objects it is necessary to make such a correction to the observed shifts of their spectral lines.

Radial velocities of nearby stars

For nearby stars, radial velocities are with very few exceptions rather small. For stars closer than 17 light-years, radial velocities range from −119 km/sec to +245 km/sec. Most values are on the order of ±20 km/sec, with a mean value of −6 km/sec.

Space motions. Space motions comprise a three-dimensional determination of stellar motion. They may be divided into a set of components related to directions in the Galaxy: *U*, directed away from the galactic centre; *V*, in the direction of galactic rotation; and *W*, toward the north galactic pole. For the nearby stars the average values for these galactic components are as follows: $U = -8$ km/sec, $V = -28$ km/sec, and $W = -12$ km/sec. These values are fairly similar to those for the galactic circular velocity components, which give $U = -9$ km/sec, $V = -12$ km/sec, and $W = -7$ km/sec. Note that the largest difference between these two sets of values is for the average *V*, which shows an excess of 16 km/sec for the nearby stars as compared to the circular velocity. Since *V* is the velocity in the direction of galactic rotation, this can be understood as resulting from the presence of stars in the local neighbourhood that have significantly elliptical orbits for which the apparent velocity in this direction is much less than the circular velocity. This fact was noted long before the kinematics of the Galaxy was understood and is referred to as the asymmetry of stellar motion.

Stream motion

The average components of the velocities of the local stellar neighbourhood also can be used to demonstrate the so-called stream motion. Calculations based on van de Kamp's table of stars within 17 light-years, excluding the star of greatest anomalous velocity, reveal that dispersions in the *V* direction and the *W* direction are approximately half the size of the dispersion in the *U* direction. This is an indication of a commonality of motion for the nearby stars; *i.e.*, these stars are not moving entirely at random but show a preferential direction of motion—the stream motion—confined somewhat to the galactic plane and to the direction of galactic rotation.

High-velocity stars. One of the nearest 45 stars, called Kapteyn's star, is an example of the high-velocity stars that lie near the Sun. Its observed radial velocity is −245 km/sec, and the components of its space velocity are $U = 19$ km/sec, $V = -288$ km/sec, $W = -52$ km/sec. The very large value for *V* indicates that with respect to circular velocity this star has practically no motion in the direction of galactic rotation at all. As the Sun's motion in its orbit around the Galaxy is estimated to be approximately 250 km/sec in this direction, the value *V* of −288 km/sec is primarily just a reflection of the solar orbital motion.

Solar motion. Solar motion is defined as the calculated motion of the Sun with respect to a specified reference frame. In practice, calculations of solar motion provide information not only on the Sun's motion with respect to its neighbours in the Galaxy but also on the kinematic properties of various kinds of stars within the system. These properties in turn can be used to deduce information on the dynamical history of the Galaxy and of its stellar components. Because accurate space motions can be obtained only for individual stars in the immediate vicinity of the Sun (within about 100 light-years), solutions for solar motion involving many stars of a given class are the prime source of information on the patterns of motion for that class. Furthermore, astronomers obtain information on the large-scale motions of galaxies in the neighbourhood of the Galaxy from solar motion solutions because it is necessary to know the space motion of the Sun with respect to the centre of the Galaxy (its orbital motion) before such velocities can be calculated.

The Sun's motion can be calculated by reference to any of three stellar motion elements: (1) the radial velocities of stars, (2) the proper motions of stars, or (3) the space motions of stars.

Solutions for solar motion based on radial velocities

Solar motion calculations from radial velocities. For objects beyond the immediate neighbourhood of the Sun, only radial velocities can be measured. Initially it is necessary to choose a standard of rest (the reference frame) from which the solar motion is to be calculated. This is usually done by selecting a particular kind of star or a portion of space. To solve for solar motion, two assumptions are made. The first is that the stars that form the standard of rest are symmetrically distributed over the sky, and the second is that the peculiar motions—the motions of individual stars with respect to that standard of rest—are randomly distributed. Considering the geometry then provides a mathematical solution for the motion of the Sun through the average rest frame of the stars being considered.

In astronomical literature where solar motion solutions are published, there is often employed a "K-term," a term that is added to the equations to account for systematic errors, the stream motions of stars, or the expansion or contraction of the member stars of the reference frame. Recent determinations of solar motion from high-dispersion radial velocities have suggested that most previous K-terms (which averaged a few kilometres per second) were the result of systematic errors in stellar spectra caused by blends of spectral lines. Of course, the K-term that arises when a solution for solar motions is calculated for galaxies results from the expansion of the system of galaxies and is very large if galaxies at great distances from the Milky Way Galaxy are included.

Solar motion calculations from proper motions. Solutions for solar motion based on the proper motions of the stars in proper motion catalogs can be carried out even when the distances are not known and the radial velocities are not given. It is necessary to consider groups of stars of limited dispersion in distance so as to have a well-defined and reasonably spatially-uniform reference frame. This can be accomplished by limiting the selection of stars according to their apparent magnitudes. The procedure is the same as the above except that the proper motion components are used instead of the radial velocities. The average distance of the stars of the reference frame enters into the solution of these equations and is related to the term often referred to as the secular parallax. The secular parallax is defined as $0.24h/r$, where h is the solar motion in astronomical units per year and r is the mean distance for the solar motion solution.

Solar motion calculations from space motions. For nearby well-observed stars, it is possible to determine complete space motions and to use these for calculating the solar motion. One must have six quantities: α (the right ascension of the star); δ (the declination of the star); μ_α (the proper motion in right ascension); μ_δ (the proper motion in declination); ρ (the radial velocity as reduced to the Sun); and r (the distance of the star). To find the solar motion, one calculates the velocity components of each star of the sample and the averages of all of these.

Solar motion solutions give values for the Sun's motion in terms of velocity components, which are normally reduced to a single velocity and a direction. The direction in which the Sun is apparently moving with respect to the reference frame is called the apex of solar motion. In addition, the calculation of the solar motion provides dispersion in velocity. Such dispersions are as intrinsically interesting as the solar motions themselves because a dispersion is an indication of the integrity of the selection of stars used as a reference frame and of its uniformity of kinematic properties. It is found, for example, that dispersions are very small for certain kinds of stars (*e.g.*, A-type stars, all of which apparently have nearly similar, almost circular orbits in the Galaxy) and are very large for some other kinds of objects (*e.g.*, the RR Lyrae variables, which show a dispersion of almost 100 km/sec due to the wide variation in the shapes and orientations of orbits for these stars).

The standard solar motion

Solar motion solutions. The motion of the Sun with respect to the nearest common stars is of primary interest. If stars within about 80 light-years of the Sun are used exclusively, the result is often called the standard solar motion. This average, taken for all kinds of stars, leads to a velocity $V_{\odot} = 19.5$ km/sec. The apex of this solar motion is in the direction of $\alpha = 270°$, $\delta = +30°$. The exact values depend on the selection of data and method of solution. These values suggest that the Sun's motion with respect to its neighbours is moderate but certainly not zero. The velocity difference is larger than the velocity dispersions for common stars of the earlier spectral types, but it is very similar in value to the dispersion for stars of a spectral type similar to the Sun. The solar velocity for, say, G5 stars, is 10 km/sec and the dispersion is 21 km/sec. Thus the Sun's motion can be considered fairly typical for its class in its neighbourhood. The peculiar motion of the Sun is a result of its relatively large age and a somewhat noncircular orbit. It is generally true that stars of later spectral types show both greater dispersions and greater values for solar motion, and this characteristic is interpreted to be the result of a mixture of orbital properties for the later spectral types, with increasingly large numbers of stars having more highly elliptical orbits.

The term basic solar motion has been used by some astronomers to define the motion of the Sun relative to stars moving in its neighbourhood in perfectly circular orbits around the galactic centre. The basic solar motion differs from the standard solar motion because of the noncircular motion of the Sun and because of the contamination of the local population of stars by the presence of older stars in noncircular orbits within the limits of the reference frame. The most commonly quoted value for the basic solar motion is a velocity of 16.5 km/sec toward an apex with a position $\alpha = 265°$, $\delta = 25°$.

When the solutions for solar motion are determined according to the spectral class of the stars, there is a correlation between the result and the spectral class. Table 7 summarizes values obtained from various sources and illustrates this fact. The apex of the solar motion, the solar motion velocity, and its dispersion are all correlated with spectral type. Generally speaking (with the exception of the very early type stars), the solar motion velocity increases with decreasing temperature of the stars, ranging from 16 km/sec for late B-type and early A-type stars to 24 km/sec for late K-type and early M-type stars. The dispersion similarly increases from a value near 10 km/sec to a value of 22 km/sec. The reason for this is related to the dynamical history of the Galaxy and the mean age and mixture of ages for stars of the different spectral types. It is quite clear, for example, that stars of early spectral type are all young, whereas stars of late spectral type are a mixture of young and old. Connected with this is the fact that the solar motion apex shows a trend for the latitude to decrease and the longitude to increase with later spectral types.

The solar motion can be based on reference frames defined by various kinds of stars and clusters of astrophysical interest. Data of this sort are interesting because of the way in which they make it possible to distinguish between objects with different kinematic properties in the Galaxy. For example, it is clear that interstellar calcium lines have relatively small solar motion and extremely small dispersion because they are primarily connected with the dust that is limited to the galactic plane and with objects that are decidedly of the Population I class. On the other hand, RR Lyrae variables and globular clusters have very large values of solar motion and very large dispersions, indicating that they are extreme Population II objects that do not all equally share in the rotational motion of the Galaxy. The solar motion of these various objects is an important consideration in determining to what population the objects belong and what their kinematic history has been.

Table 7: Adopted Components of the Solar Motion and Velocity Dispersions

type	solar motion (km/sec)			spread in velocities (km/sec)		
	U_s	V_s	W_s	U	V	W
cO-cB5	−9.0	+13.4	+3.7	12	11	9
cF-cM	−7.9	+11.7	+6.5	13	9	7
gA	−13.4	+11.6	+10.3	22	13	9
gF	−19.7	+18.5	+9.5	28	15	9
gG	−7.2	+11.1	+6.9	26	18	15
gK0	−10.6	+18.6	+6.5	31	21	16
gK3	−9.0	+17.6	+6.4	31	21	17
gM	−4.5	+18.3	+6.2	31	23	16
Carbon stars	−10.7	+31.8	+3.5	48	23	16
Subgiants	−8.0	+28.0	+8.0	43	27	24
B0	−9.6	+14.5	+6.7	10	9	6
dA0	−7.3	+13.7	+7.2	15	9	9
dA5	−8.5	+7.8	+7.4	20	9	9
dF5	−10.1	+12.3	+6.2	27	17	17
dG0	−14.5	+21.1	+6.4	26	18	20
dG5	−8.1	+22.1	+4.3	32	17	15
dK0	−10.8	+14.9	+7.4	28	16	11
dK5	−9.5	+22.4	+5.8	35	20	16
dM0	−6.1	+14.6	+6.9	32	21	19
dM5	−9.8	+19.3	+8.6	31	23	16
White dwarfs	−6	+37	+8	50	33	25
Planetary nebulas	−8	+29	+8	45	35	20
Classical Cepheids	−8.6	+12.0	+7.6	13	9	5
Interstellar Ca II	−11.4	+14.4	+8.2	...	6	...

When some of these classes of objects are examined in greater detail, it is possible to separate them into subgroups and find correlations with other astrophysical properties. Take, for example, globular clusters, for which the solar motion is correlated with the spectral type of the clusters. The clusters of spectral types G0–G5 (the more metal-rich clusters) have a mean solar motion of 80 ± 82 km/sec (corrected for the standard solar motion). The earlier type globular clusters of types F2–F9, on the other hand, have a mean velocity of 162 ± 36 km/sec, suggesting that they partake much less extensively in the general rotation of the Galaxy. Similarly, the most distant globular clusters have a larger solar motion than the ones closer to the galactic centre. Studies of RR Lyrae variables also show correlations of this sort. The period of an RR Lyrae variable, for example, is correlated with its motion with respect to the Sun. For type ab RR Lyrae variables, periods frequently vary from 0.3 to 0.7 day, and the range of solar motion for this range of period extends from 30 to 205 km/sec, respectively. This condition is believed to be primarily the result of the effects of the spread in age and composition for the RR Lyrae variables in the field, which is similar to but larger than the spread in the properties of the globular clusters.

The solar velocity of rotation

Since the direction of the centre of the Galaxy is well established by radio measurements and since the galactic plane is clearly established by both radio and optical studies, it is possible to determine the motion of the Sun with respect to a fixed frame of reference centred at the Galaxy and not rotating (*i.e.*, tied to the external galaxies). The value for this motion is generally accepted to be 225 km/sec in the direction $\ell^{\text{II}} = 90°$. It is not a firmly established number, but it is used by convention in most studies.

In order to arrive at a clear idea of the Sun's motion in the Galaxy as well as the motion of the Galaxy with respect to neighbouring systems, solar motion has been studied with respect to the Local Group galaxies and those in nearby space. Hubble determined the Sun's motion with respect to the galaxies beyond the Local Group and found the value of 300 km/sec in the direction toward galactic longitude 120°, latitude +35°. This velocity includes the Sun's motion in relation to its proper circular velocity, its circular velocity around the galactic centre, the motion of the Galaxy with respect to the Local Group, and the latter's motion with respect to its neighbours.

The magnetic field of the Galaxy. It was once thought that the spiral structure of galaxies might be controlled by a strong magnetic field. However, when the general magnetic field was detected by radio techniques, it was found to be too weak to have large-scale effects on galactic structure. The strength of the galactic field is only about 0.000001 times the strength of the Earth's field at its surface, a value that is much too low to have dynamical effects on the interstellar gas that could account for the order represented by the spiral-arm structure. This is, however, sufficient strength to cause a general alignment of the dust grains in interstellar space, a feature that is detected by measurements of the polarization of starlight. In the prevailing model of interstellar dust grains, the particles are shown to be rapidly spinning and to contain small amounts of metal (probably iron), though the primary constituents are ice and carbon. The magnetic field of the Galaxy can gradually act on the dust particles and cause their rotational axes to line up in such a way that their short axes are parallel to the direction of the field. The field itself is aligned along the Milky Way band, so that the short axes of the particles also become aligned along the galactic plane. Polarization measurements of stars at low galactic latitudes confirm this pattern.

The rotation of the Galaxy. As discussed above, the motions of stars in the local stellar neighbourhood can be understood in terms of a general population of stars that have circular orbits of rotation around the distant galactic nucleus, with an admixture of stars that have more highly elliptical orbits and that appear to be high-velocity stars to a terrestrial observer as the Earth moves with the Sun in its circular orbit. The general rotation of the disk stars was first detected through studies made in the 1920s, notably those of the Swedish astronomer Bertil Lindblad, who correctly interpreted the apparent asymmetries in stellar motions as the result of this multiple nature of stellar orbital characteristics.

Differential rotation and its effect

The disk component of the Galaxy rotates around the nucleus in a manner similar to the pattern for the planets of the solar system, which have nearly circular orbits around the Sun. Because the rotation rate is different at different distances from the centre of the Galaxy, the measured velocities of disk stars in different directions along the Milky Way exhibit different patterns. The Dutch astronomer Jan H. Oort first interpreted this effect in terms of galactic rotation motions, employing the radial velocities and proper motions of stars. He demonstrated that differential rotation leads to a systematic variation of the radial velocities of stars with galactic longitude following the mathematical expression:

$$\text{radial velocity} = Ar \sin 2l,$$

where A is called Oort's constant and is approximately 15 km/sec/kpc, r is the distance to the star, and l is the galactic longtitude (Figure 9).

A similar expression can be derived for measured proper motions of stars. The agreement of observed data with Oort's formulas was a landmark demonstration of the correctness of Lindblad's ideas about stellar motions. It led to the modern understanding of the Galaxy as a giant rotating disk.

From van de Kamp in B J Bok and P F Bok, *The Milky Way* (1957), Harvard University Press

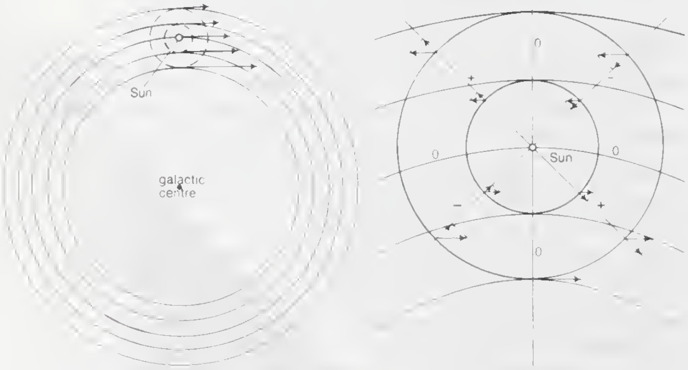


Figure 9: The effect of galactic rotation on the radial velocities (speeds in the line of sight). (Left) The arrows represent the variation of rotational velocity with the distance from the centre. (Right) The arrows represent the velocities as observed. The radial velocity components are also shown (see text).

The external galaxies

THE EXTRAGALACTIC DISTANCE SCALE

Before astronomers could establish the existence of galaxies, they had to develop a way to measure their distances. In an earlier section, it was explained how astronomers first accomplished this exceedingly difficult task for the nearby galaxies during the 1920s. Since that time, progress has been slow and the results far from satisfactory. Distance determinations for the nearest galaxies still remain uncertain by as much as 10 percent, and the scale of distances beyond the Local Group of galaxies is even more unsure, with an uncertainty of possibly a factor of two. The reason for this unfortunate situation is the great difficulty of observing distant galaxies in sufficient detail to recognize the necessary distance criteria accurately, to measure their characteristics, and to make certain that their characteristics are correctly interpreted.

From Paul W Hodge, *The Physics and Astronomy of Galaxies and Cosmology* (1966)

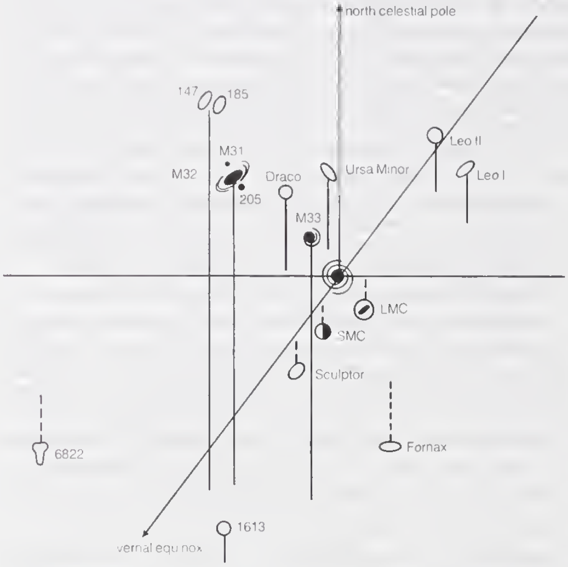


Figure 10: Three-dimensional schematic diagram of the principal galaxies of the Local Group. Galaxy images are not to scale.

The process involved is one of many successive steps that are all closely tied to one another. Distances are first determined for a number of galaxies close to the Milky Way Galaxy, specifically those in the Local Group and a few in other nearby groups. For this step, criteria are used that have been calibrated within the Galaxy, where checks can be made between different methods and where the ultimate criterion is a geometrical one (basically involving trigonometric parallaxes and the moving-cluster method). Next, the nearby galaxies are used to set up new, brighter distance criteria that can be employed in more distant realms where only the brightest stars and other objects are discernible in galaxies. These in turn can be used to establish criteria based on the properties of whole galaxies so that distances can be found for systems so far away that no individual objects can be resolved within them. These last criteria then are applied to extend scientific knowledge to the most distant galaxies that can be detected—those at the very edge of the visible universe. Considering the magnitude of the task, it is no surprise that the answers obtained are not yet highly certain. Astronomers have done remarkably well to measure such vast distances even with an uncertainty of a factor of two.

The Local Group of galaxies is a concentration of approximately 35 galaxies, dominated by two large spirals, the Milky Way system and the Andromeda galaxy (Table 8 and Figure 10). For many of these galaxies, distances can be measured using the Cepheid period–luminosity law, which has been refined and made more accurate since Hubble first used it. For instance, the nearest ex-

Table 8: The Members of the Local Group

name	Hubble type	absolute visual magnitude	name	Hubble type	absolute visual magnitude
M31	Sb	−21.1	And II	E	−11.8
Milky Way	Sb/c	−20.6	Leo I	E3	−11.7
M33	Sc	−18.9	DDO 210	Irr	−11.5
LMC	Irr	−18.1	GR8	Irr	−11.4
M32	E2	−16.4	SagDIG	Irr	−11.2
NGC 6822	Irr	−16.4	Sculptor	E3	−10.7
NGC 205	E5	−16.3	And III	E	−10.3
SMC	Irr	−16.2	LGS 3	Irr	−10.2
NGC 185	E3	−15.3	Sextans I	E	−10.0
NGC 147	E5	−15.1	Phoenix	Irr	−9.9
IC 1613	Irr	−14.9	Tucana	E	−9.5
WLM	Irr	−14.1	Leo II	E0	−9.4
Leo A	Irr	−14.0	Ursa Minor	E5	−8.9
Pegasus	Irr	−13.9	Draco	E3	−8.6
Fornax	E3	−13.7	Carina	E4	−7.6
UGC A86	Irr	−12.7	EGB0427	Irr	?
And I	E	−11.8			

ternal galaxy, the Large Magellanic Cloud, contains more than 2,000 Cepheid variables, which can be compared to Cepheids of known distance in the Galaxy to yield a distance determination of 150,000 light-years. This method has been employed for 10 galaxies of the Local Group (Figure 11). Most of the rest of the members are elliptical galaxies, which do not have Cepheid variables; their distances are measured by using Population II stars, such as RR Lyrae variables.

Beyond the Local Group are two nearby groups for which the period–luminosity relation has been used: the Sculptor group and the M81 group. Both of these are small clusters of galaxies that are similar in size to the Local Group. They lie at a distance of from 10,000,000 to 15,000,000 light-years.

Distance-determining techniques

While the period–luminosity relation is the primary distance criterion in this realm, others also are used. One of these is main-sequence fitting, a technique that compares the temperatures of individual stars (those in their normal, stable main-sequence phase) with similar stars in steller clusters in the Galaxy. Main-sequence fitting became effective in the 1980s, when powerful new detectors allowed accurate measurements of very faint stars in the few nearest galaxies.

Also in the 1980s, astronomers developed a new method of effectively measuring distances to galaxies in the Local Group, in nearby groups, and even as far away as the Virgo cluster, lying at a distance of about 50,000,000 light-years. This method makes use of planetary nebulas, the ringlike shells that surround some stars in their late stages of evolution. Planetary nebulas have a variety of luminosities, depending on their age and other physical circumstances; however, it has been determined that the brightest planetary nebulas have an upper limit to their intrinsic brightnesses. This means that astronomers can measure the brightnesses of such nebulas in any given galaxy, find the upper limit to the apparent brightnesses, and then immediately calculate the distance of the galaxy. Another recent method involves the use of the nova phenomenon, which can be detected in galaxies in the Local Group and beyond—that is to say, as remote as the Virgo cluster.

From Paul W. Hodge, *The Physics and Astronomy of Galaxies and Cosmology* (1966)

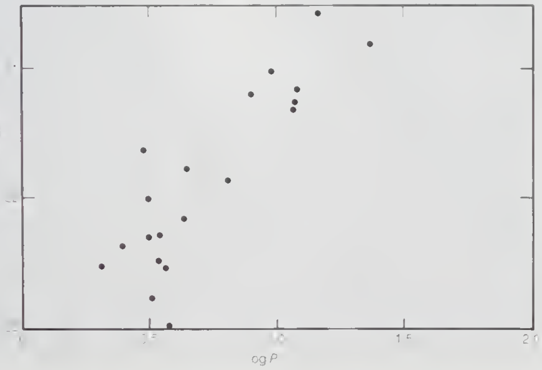


Figure 11: Period–luminosity relation for Cepheids in a portion of M31. Open circles are type II Cepheids.

Once distances have been established for these nearby galaxies and groups, new criteria are calibrated for extension to fainter galaxies. Examples of the many different criteria that have been tried are the luminosities of the brightest stars in the Galaxy, the diameters of the largest H II regions, supernova luminosities, the spread in the rotational velocities of stars and interstellar gas, and the luminosities of globular clusters. All of these criteria have difficulties in their application because of dependencies on galaxy type, composition, luminosity, and other characteristics, so that the results of several methods must be intercompared and cross-checked. Such distance criteria allow astronomers to measure the distances to galaxies out to a few hundred million light-years.

The Hubble constant

Beyond 100,000,000 light-years, another method becomes possible. The expansion of the universe, at least for the immediate neighbourhood of the Local Group (within 1,000,000,000 light-years or so), is linear enough that the radial velocity of a galaxy is à reliable distance indicator. The velocity is directly proportional to the distance in this interval, so that once a galaxy's radial velocity has been measured all that must be known is the constant of proportionality, which is called the Hubble constant. Although there still remains some uncertainty in the correct value of the Hubble constant, research in the 1990s has indicated that the constant has a value very near 25 km/sec per 1,000,000 light-years (see Figure 12). Radial velocity for nearby galaxies and groups is affected by the Local Group's motion with respect to the general background of galaxies toward a concentration of galaxies and groups of galaxies centred on the Virgo cluster, all of which make up a local supercluster. Radial velocities cannot give reliable distances beyond a few billion light-years because, in the case of such galaxies, astronomers are looking so far back in time that they do not know whether the expansion rate of the universe was then the same as it is now. The light that is observed today was emitted several billion years ago when the universe was much younger and smaller than it is at present.

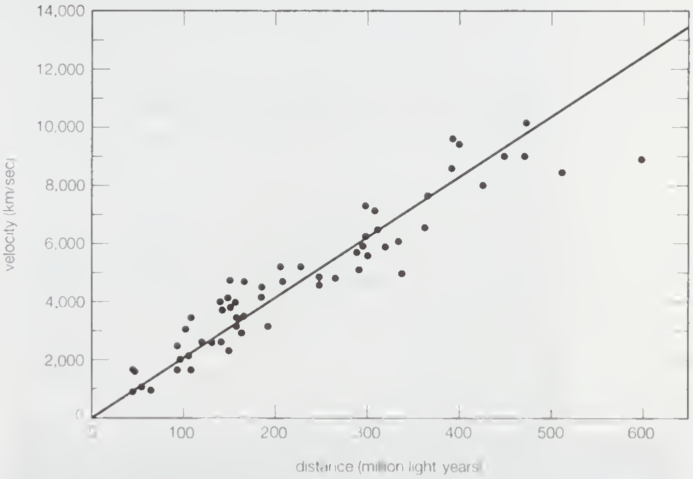


Figure 12: Hubble diagram for the brightest galaxies in clusters. The line is a fit that corresponds to a Hubble constant of 25 km/sec per 1,000,000 light-years.

To find the distances of very distant galaxies, astronomers have to avail themselves of methods that make use of the total properties of the very brightest galaxies. Most commonly it is assumed that the brightest galaxies in clusters all have the same true luminosity (this appears to be the case for relatively nearby clusters for which distances can be measured in other ways) and that therefore measuring the apparent brightness of the brightest galaxy in a distant cluster will give its distance. This method, as well as others that make use of the cluster environment, have yielded distance estimates as large as 10,000,000,000 to 20,000,000,000 light-years for the most distant galaxies detectable (Figure 13).

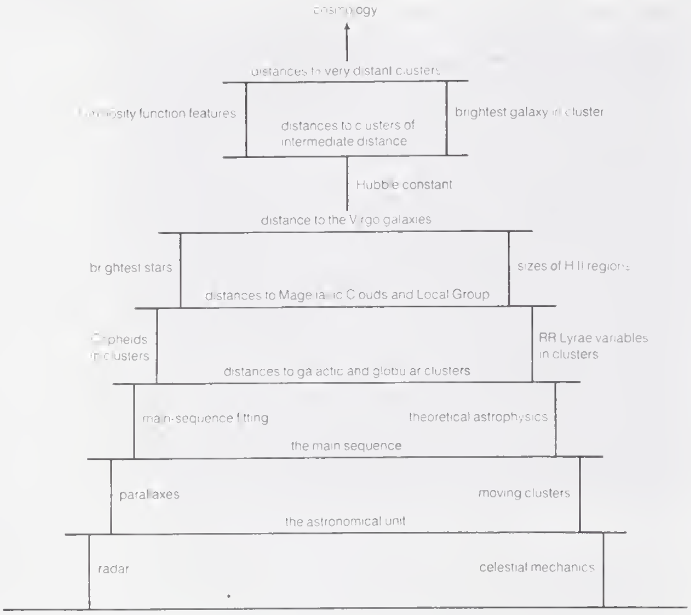

cosmology
distances to very distant clusters
luminosity function features
brightest galaxy in cluster
distances to clusters of intermediate distance
Hubble constant
distance to the Virgo galaxies
brightest stars
sizes of H II regions
distances to Magellanic Clouds and Local Group
Cepheids in clusters
RR Lyrae variables in clusters
distances to galactic and globular clusters
main-sequence fitting
theoretical astrophysics
the main sequence
parallaxes
moving clusters
the astronomical unit
radar
celestial mechanics

Figure 13: Framework of extragalactic distance-scale determinations.
From Paul W. Hodge, *The Physics and Astronomy of Galaxies and Cosmology* (1966)

PHYSICAL PROPERTIES OF EXTERNAL GALAXIES

Size and mass. The range in intrinsic size for the external galaxies extends from the smallest systems, such as the nearby dwarf galaxy GR8 with a diameter of about 5,000 light-years, to giant radio galaxies, the extent of which (including their radio-bright lobes) is more than 3,000,000 light-years. Normal large spiral galaxies, such as the Andromeda galaxy, have diameters of 100,000 to 500,000 light-years.

The total masses of galaxies are not well known, largely because of the uncertain nature of the hypothesized invisible dark halos that surround many, or possibly all, galaxies. The total mass of material within the radius out to which the light of a galaxy can be detected is known for many hundreds of systems. The range is from about 100,000 to roughly 1,000,000,000,000 times the Sun's mass. The mass of a typical large spiral is about 500,000,-000,000 suns.

Luminosity. The external galaxies show an extremely large range in their total luminosities. The intrinsically faintest are the extreme dwarf elliptical galaxies, such as the Ursa Minor dwarf, which has a luminosity of approximately 100,000 suns. The most luminous galaxies are those that contain quasars at their centres. These remarkably bright, superactive nuclei can be as luminous as 2,000,000,000,000 suns. The underlying galaxies are often as much as 100 times fainter than their nuclei. Normal large spiral galaxies have a luminosity of a few hundred billion suns.

Super-
active
nuclei

Age. Even though different galaxies have had quite different histories, measurements tend to suggest that most, if not all, galaxies have very nearly the same age. The age of the Milky Way Galaxy, which is measured by determining the ages of the oldest stars found within it, is approximately 16,000,000,000 years. Nearby galaxies, even those such as the Large and Small Magellanic Clouds that contain a multitude of very young stars, also have at least a few very old stars of approximately that same age. When more distant galaxies are examined, their spectra and colours closely resemble those of the nearby galaxies, and it is inferred that they too must contain a population of similarly very old stars. Extremely distant galaxies, on the other hand, look younger, but that is because the "look-back" time for them is a significant fraction of their age; the light received from such galaxies was emitted when they were appreciably younger.

It seems likely that all of the galaxies began to form at about the same time when the universe had cooled down enough for matter to condense, and thus they all started forming stars during nearly the same epoch. Their large differences are not a matter of age but rather of how they proceeded to regulate the processing of their materials (gas and dust) into stars. The ellipticals formed almost all their stars during the first few billion years, while the spirals and the irregulars have been using up their materials more gradually.

Composition. The abundances of the chemical elements in stars and galaxies are remarkably uniform. The ratios of the amounts of the different elements that astronomers observe for the Sun is a reasonably good approximation for those of other stars in the Galaxy and also for stars in other galaxies. The main difference found is in the relative amount of the primordial gases, hydrogen and helium. As noted earlier, the heavier elements are formed by stellar evolutionary processes, and they are relatively more abundant in areas where extensive star formation has been taking place. Thus, in such small elliptical galaxies as the Draco system where almost all of the stars were formed at the beginning of its lifetime, the component stars are nearly pure hydrogen and helium, while in such large galaxies as the Andromeda Galaxy, there are areas where star formation has been active for a long time (right up to the present in fact), and here investigators find that the heavier elements are more abundant. In some external galaxies as well as in some parts of the Milky Way system, heavy elements are even more abundant than in the Sun but rarely by more than a factor of two or so. Even in such cases, hydrogen and helium make up most of the constituent materials, accounting for at least 90 percent of the mass.

Relative
amount of
hydrogen
and helium

STRUCTURE

The spheroidal component. Most and perhaps all galaxies have a spheroidal component of very old stars. In the ellipticals this component comprises all or most of any given system. In the spirals it represents about half of the constituent stars (this fraction varies greatly according to galaxy type). In the irregulars the spheroidal component is very inconspicuous or, possibly in a few cases, entirely absent. The structure of the spheroidal component of all galaxies is similar, as if the spirals and irregulars possess a skeleton of old stars arranged in a structure that resem-

By courtesy of the National Optical Astronomy Observatories

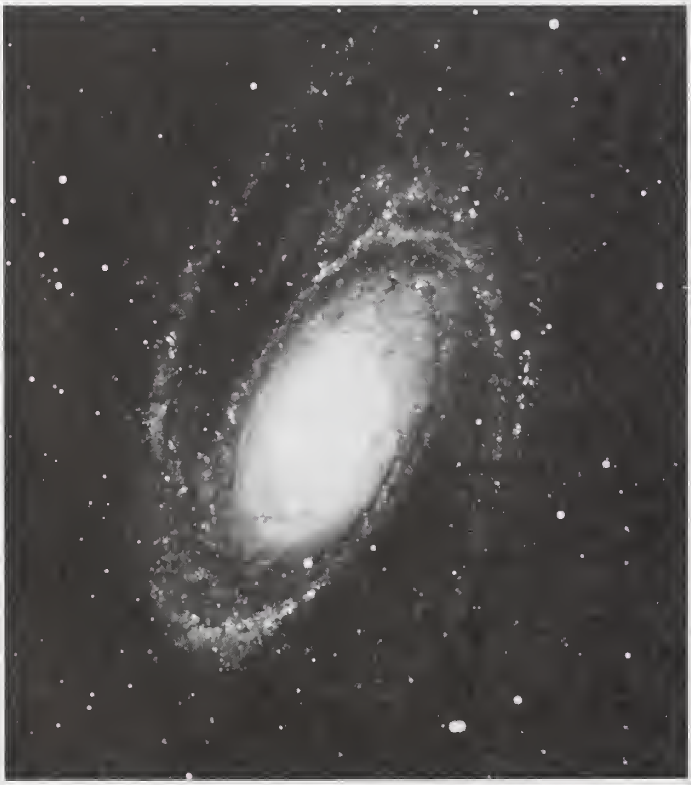


Figure 14: M81, a large spiral galaxy in the constellation Ursa Major.

bles an elliptical. The radial distribution of stars follows a law of the form

$$I = I_e 10^{(-3.33([r/r_e]^{1/4}-1))},$$

where I is the surface brightness (or the stellar density) at position r, r is the radial distance from the centre, and I_e and r_e are constants. This expression, advocated by de Vaucouleurs, is an empirical formula that works remarkably well in describing the spheroidal components of almost all galaxies. An alternative formula, put forth by Hubble, is of the form

$$I = I_0(r/a + 1)^{-2},$$

where I is the surface brightness, I_0 is the central brightness, r is the distance from the centre, and a is a scaling constant. Either of these formulas describes the structure well, but neither explains it.

A somewhat more complicated set of equations can be derived on the basis of the mutual gravitational attraction of stars for one another and the long-term effects of close encounters between stars. These models of the spheroidal component (appropriately modified in the presence of other galactic components) fit the observed structures well. Rotation is not an important factor, since elliptical galaxies and the spheroidal component of spiral systems (*e.g.*, the Milky Way Galaxy) rotate slowly. One of the open questions about the structure of these objects is why they have as much flattening as some of them do. In most cases, the measured rotation rate is inadequate to explain the flattening on the basis of a model of an oblate spheroid that rotates around its short axis. It has been suggested that some elliptical galaxies may instead be prolate spheroids that rotate around their long axis.

Slow rotation rate

The disk component. Except for such early type galaxies as S0, SB0, Sa, and SBa systems, spirals and irregulars have a flat component of stars that emits most of their brightness. The disk component has a thickness that is approximately one-fifth its diameter (this varies, depending on the type of stars being considered, as explained above for the Galaxy). The stars show a radial distribution that obeys an exponential decrease outward; *i.e.*, the brightness obeys a formula of the form

$$\log I = -kr,$$

where I is the surface brightness, r is the distance from the centre, and k is a scaling constant. This constant is dependent both on the type of the galaxy and on its intrinsic luminosity. The steepness of the outward slope is greatest for the early Hubble types (Sa and SBa) and for the least luminous galaxies.

Spiral arms. The structure of the arms of spiral galaxies depends on the galaxy type, and there is also a great deal of variability within each type. Generally the early Hubble types have smooth, indistinct spiral arms with small pitch angles. The later types have more open arms (larger pitch angles; Figure 14). Within a given type there can be found galaxies that have extensive arms (extending around the centre for two or more complete rotations) and those that have a chaotic arm structure made up of many short fragments that extend only 20° or 30° around the centre. All spiral arms fit reasonably well to a logarithmic spiral of the form described above for the Milky Way Galaxy.

Gas distribution. If one were to look at galaxies at wavelengths that show only neutral hydrogen gas, they would look rather different from their optical appearance. Normally the gas, as detected at radio wavelengths for neutral hydrogen atoms, is more widely spread out, with the size of the gas component often extending to twice the size of the optically visible image. Also, in some galaxies a hole exists in the centre of the system where almost no neutral hydrogen occurs. There is, however, enough molecular hydrogen to make up for the lack of atomic hydrogen. Molecular hydrogen is difficult to detect, but it is accompanied by other molecules, such as carbon monoxide, which can be observed at radio wavelengths.

Neutral hydrogen

CLUSTERS OF GALAXIES

Galaxies tend to cluster together, sometimes in small groups and sometimes in enormous complexes. Most galaxies have companions, either a few nearby objects or a large-scale cluster; isolated galaxies, in other words, are quite rare (Table 9).

Table 9: Properties of Some Well-Known Clusters of Galaxies

name	distance	diameter	description
	(millions of light-years)		
Local group	—	3	small loose group
M81 group	7	1	loose group
Virgo cluster	50	7	large complex cluster
Fornax cluster	50	3	complex cluster
Coma cluster	150	23	large compact cluster
Hercules cluster	240	7	large loose cluster
Corona Borealis cluster	700	7	compact cluster

Types of clusters. There are several different classification schemes for galaxy clusters, but the simplest is the most useful. This scheme divides clusters into three classes: groups, irregulars, and sphericals.

Groups. The groups class is composed of small, compact groups of 10 to 50 galaxies of mixed types, spanning roughly 5,000,000 light-years. An example of such an entity is the Local Group, which includes the Milky Way Galaxy, the Magellanic Clouds, the Andromeda Galaxy, and 18 other systems mostly of the dwarf variety.

Irregular clusters. Irregular clusters are large, loosely structured assemblages of mixed galaxy types (mostly spirals and ellipticals), totaling perhaps 1,000 or more systems and extending out 10,000,000 to 50,000,000 light-years. The Virgo and Hercules clusters are representative of this class (Figure 15).

Spherical clusters. Spherical clusters are dense and consist almost exclusively of elliptical and S0 galaxies. They are enormous, having a linear diameter of up to 50,000,000 light-years. Spherical clusters may contain as many as 10,000 galaxies, which are concentrated toward the cluster centre.

Distribution. Clusters of galaxies are found all over the sky. They are difficult to detect along the Milky Way,

By courtesy of the National Optical Astronomy Observatories

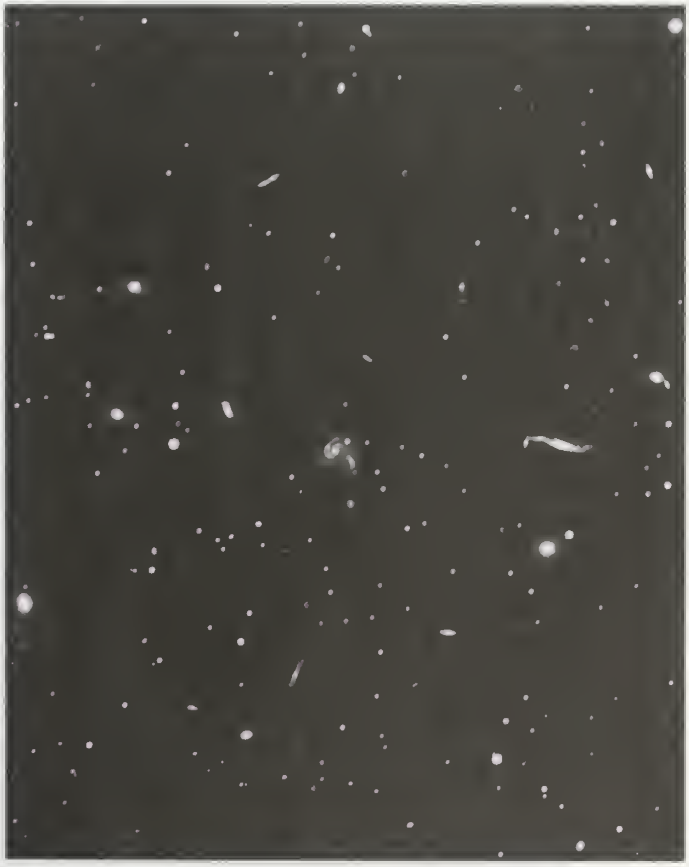


Figure 15: Hercules cluster of galaxies, a vast, loosely structured assemblage composed chiefly of spirals and ellipticals.

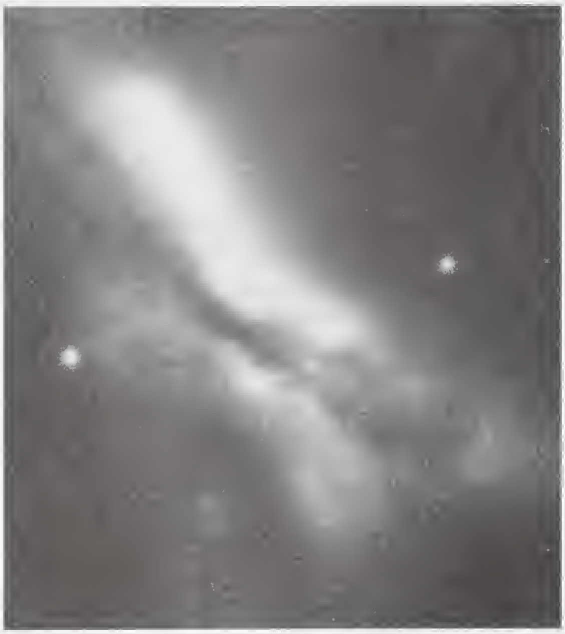


Figure 16: NGC 520, an example of an interacting pair of galaxies, in which the shapes of the galaxies and the motions of the constituent stars and interstellar material are grossly distorted.

Lick Observatory, Mount Hamilton, California

where high concentrations of the Galaxy's dust and gas obscure virtually everything at optical wavelengths. However, even there clusters can be found in a few galactic "windows," the random holes in the dust that permit optical observations.

Superclusters

The clusters are not evenly spaced in the sky; instead, they are arranged in a way that suggests a certain amount of organization. Clusters are frequently associated with other clusters, forming giant superclusters. These superclusters typically consist of three to 10 clusters and span as many as 200,000,000 light-years. There also are immense areas between clusters that are nearly empty, forming "voids." Large-scale surveys made in the 1980s of the radial velocities of galaxies revealed an even larger kind of structure. It was discovered that galaxies and galaxy clusters tend to fall in position along large planes and curves, almost like giant walls, with relatively empty spaces between them. A related large-scale structure was found to exist where there occur departures from the velocity–distance relation in certain directions, indicating that the otherwise uniform expansion is being perturbed by large concentrations of mass. One of these, discovered in 1988, has been dubbed "The Great Attractor."

Interactions between cluster members. Galaxies in clusters exist in a part of the universe that is much denser than average, and the result is that they have several unusual features. In the inner parts of dense clusters there are very few, if any, normal spiral galaxies. This condition is probably the result of fairly frequent collisions between the closely packed galaxies, as such violent interactions tend to sweep out the interstellar gas, leaving behind only the spherical component and a gasless disk. What remains, in effect, is an S0 galaxy.

A second and related effect of galaxy interactions is the presence of gas-poor spiral systems at the centres of large irregular clusters. A significant number of the members of such clusters have anomalously small amounts of neutral hydrogen, and their gas components are smaller on the average than those for more isolated galaxies. This is thought to be the result of frequent distant encounters between such galaxies involving the disruption of their outer parts (Figure 16).

A third effect of the dense cluster environment is the presence in some clusters—usually rather small, dense clusters—of an unusual type of galaxy called a cD galaxy. These objects are somewhat similar in structure to S0 galaxies, but they are considerably larger, having envelopes that extend out to radii as large as 1,000,000 light-years. Many of them have multiple nuclei, and most are strong sources of radio waves. The most likely explanation for cD galaxies is that they are massive, central galactic systems that have captured smaller cluster members because of their dominating gravitational fields and have absorbed the other galaxies into their own structures. Astronomers refer to this process as galactic cannibalism. In this sense, the outer extended disks of cD systems, as well as their multiple nuclei, represent the remains of past, partly digested "meals" (Figure 17).

Galactic cannibalism

By courtesy of the National Optical Astronomy Observatories

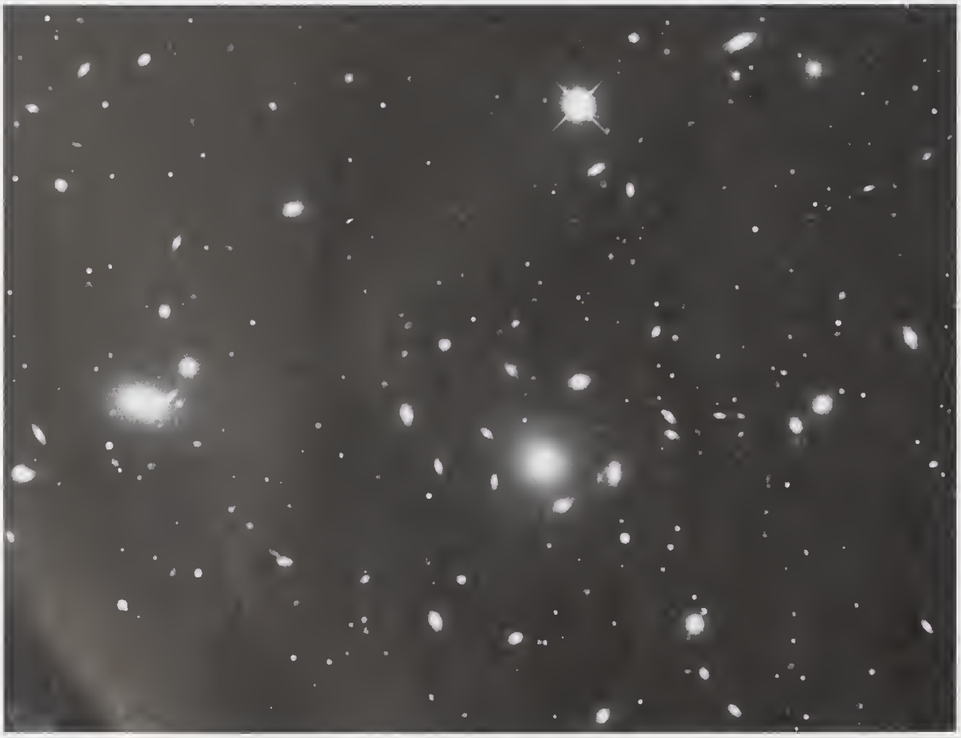


Figure 17: Coma cluster, a spherically symmetrical group of galaxies with a high concentration of ellipticals toward its centre.

One more effect that can be traced to the cluster environment is the presence of strong radio and X-ray sources, which tend to occur in or near the centres of clusters of galaxies. These will be discussed in detail in the next section.

EXTRAGALACTIC RADIO AND X-RAY SOURCES

Radio galaxies. Some of the strongest radio sources in the sky are galaxies (Table 10). Most of them have a peculiar morphology that is related to the cause of their radio radiation. Some are relatively isolated galaxies, but most galaxies that emit unusually large amounts of radio energy are found in large clusters.

The basic characteristics of radio galaxies and the variations that exist among them can be made clear with two examples. The first is Centaurus A, a giant radio structure surrounding a bright, peculiar galaxy of remarkable morphology designated NGC 5128. It exemplifies a type of radio galaxy that consists of an optical galaxy located at the centre of an immensely larger two-lobed radio source. In the particular case of Centaurus A, the extent of the radio structure is so great that it is almost 100 times the size of the central galaxy, which is itself a giant galaxy. This radio structure includes, besides the pair of far-flung radio lobes, two other sets of radio sources: one that is approximately the size of the optical galaxy and that resembles the outer structure in shape; and a second that is an intense, small source at the galaxy's nucleus (Figure 18). Optically, NGC 5128 appears as a giant elliptical galaxy with two notable characteristics: it has an unusual disk of dust and gas surrounding it and thin jets of interstellar gas and young stars radiating outward (Figure 19). The most plausible explanation currently offered for this whole array is that a series of explosive events in the nucleus of the galaxy expelled hot, ionized gas from the centre at relativistic velocities (*i.e.*, those at nearly the speed of light) in two opposite directions. These clouds of relativistic particles generate synchrotron radiation, which is detected at radio (and X-ray) wavelengths. In this model the very large structure is associated with an old event, while the inner lobes are the result of a more recent explosion. The centre is still active, as evidenced by the presence of the nuclear radio source.

Centaurus A

Virgo A

The other notable example of a radio galaxy is Virgo A, a powerful radio source that corresponds to a bright elliptical galaxy in the Virgo cluster designated as M87. In this type of radio galaxy, most of the radio radiation is emitted from an appreciably smaller area than in the case of Centaurus A. This area coincides in size with the optically visible object (Figure 20). Virgo A is not particularly unusual except for one peculiarity: it has a bright jet of gaseous material that appears to emanate from the nucleus of the galaxy, extending out approximately halfway to its faint outer parts. This gaseous jet can be detected at optical, radio, and other (*e.g.*, X-ray) wavelengths; its spectrum suggests strongly that it shines by means of the synchrotron mechanism.

About the only condition that can account for the immense amounts of energy emitted by radio galaxies is the capture of material (interstellar gas and stars) by a supermassive object at their centre. Such an object would resemble the one thought to be in the nucleus of the Milky Way Galaxy but far more massive. In short, the most probable type of supermassive object for explaining the details of strong radio sources would be a black hole. Large amounts of energy can be released when material is captured by a black hole. An extremely hot, high-density accretion disk is first formed around the supermassive object from the material, and then some of the material seems to be ejected explosively from the area, giving rise to the various radio jets and lobes observed.

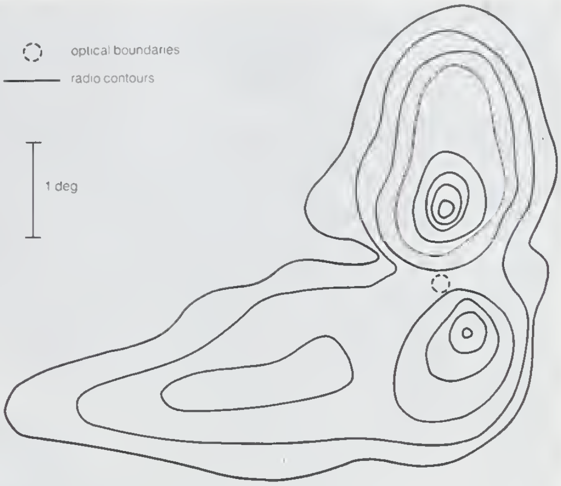


Figure 18: Continuum radio emission from the immense radio source Centaurus A (NGC 5128 is the optical galaxy at the centre of the source).

From Paul W. Hodge, *The Physics and Astronomy of Galaxies and Cosmology* (1966)

X-ray galaxies. Synchrotron radiation is characteristically emitted at virtually all wavelengths at almost the same intensity. A synchrotron source therefore ought to be detectable at optical and radio wavelengths, as well as at others (*e.g.*, infrared, ultraviolet, X-ray, and gamma-ray wavelengths). For radio galaxies this does seem to be the case, at least in circumstances where the radiation is not screened by absorbing material in the source or in intervening space.

X rays are absorbed by the Earth's atmosphere. Consequently, X-ray galaxies could not be detected until it became possible to place telescopes above the atmosphere, first with balloons and sounding rockets and later with orbiting observatories specially designed for X-ray studies. For example, the Einstein Observatory, which was in operation during the early 1980s, made a fairly complete search for X-ray sources across the sky and studied several of them in detail. Many of the sources turned out to be distant galaxies and quasars, while others were relatively nearby objects, including neutron stars (extremely dense stars composed almost exclusively of neutrons) in the Milky Way Galaxy.

A substantial number of the X-ray galaxies so far detected are also well-known radio galaxies. Some X-ray sources, such as certain radio sources, are much too large to be individual galaxies, consisting rather of a whole cluster of galaxies.

Clusters of galaxies as radio and X-ray sources. Some clusters of galaxies contain a widespread intergalactic cloud of hot gas that can be detected as a diffuse radio source or

Table 10: Some Strong Radio Sources

name	optical description	radio luminosity (ergs/sec)	radio diameter	optical diameter	distance (millions of light-years)
Virgo A (M87)	elliptical galaxy with "jet" from nucleus	5×10^{41}	10′ (and 30″ jet)	10′ (and 20″ jet)	50
Fornax A (NGC 1316)	S0 galaxy with dust lanes and outer ring	6×10^{41}	30′ (double)	15′	50
Centaurus A (NGC 5128)	peculiar spiral	8×10^{41}	10° (multiple)	20′	15
Perseus A (NGC 1275)	complex S0 galaxy	10^{42}	4′ (with 10″ core)	2′	230
Hydra A	large S0 galaxy	2×10^{43}	50″	0′5	680
Hercules A	large S0 galaxy with dust lanes and outer wisps	1.5×10^{44}	110′ (double)	0′2	2,000
Cygnus A	double galaxy	5×10^{44}	106″ (double)	2″	710
3C 48	quasi-stellar source, fluctuations in brightness	5×10^{44}	<1″	<0″5	3,600

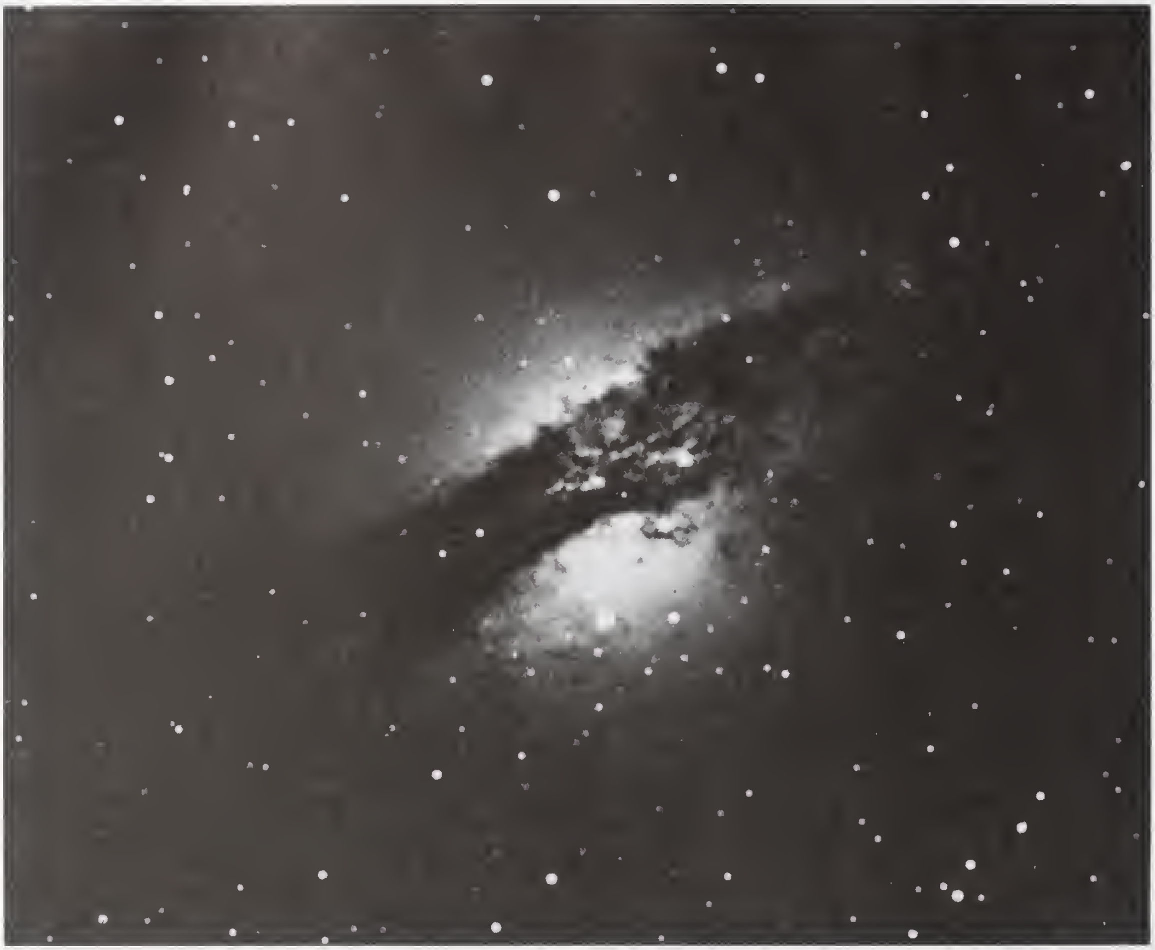


Figure 19: Centaurus A, when viewed at visible wavelengths, appears as a bright ellipse in the middle of which lies a disk-shaped absorption lane composed largely of dust and ionized atomic hydrogen.

By courtesy of the National Optical Astronomy Observatories

as a large-scale source of X rays. The gaseous cloud has a low density but a very high temperature, having been heated by the passage of the galaxies of a cluster through it and by the emission of high-energy particles from active galaxies within it.

"Head–tail" galaxies

The form of certain radio galaxies in clusters points rather strongly to the presence of intergalactic gas. These are the "head–tail" galaxies, systems that have a bright source accompanied by a tail or tails that appear swept back by their interaction with the cooler, more stationary intergalactic gas. These tails are radio lobes of ejected gas whose shape has been distorted by collisions with the cluster medium.

From Paul W Hodge, *The Physics and Astronomy of Galaxies and Cosmology* (1966)

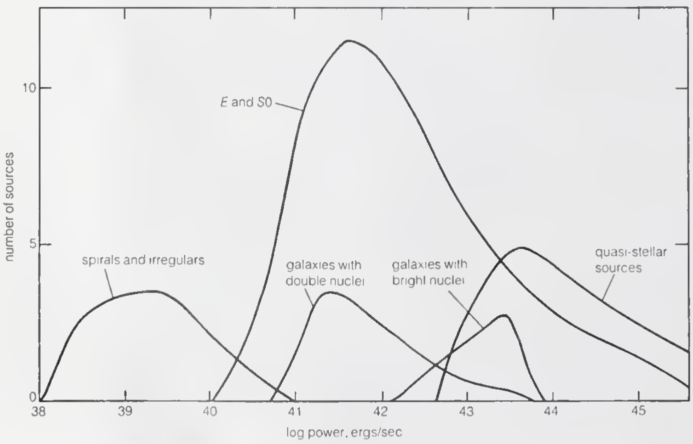


Figure 20: Distribution of radio sources according to radiated power and optical type.

Quasars. An apparently new kind of radio source was discovered in the early 1960s when radio astronomers identified a very small but powerful radio object designated 3C 48 with a stellar optical image. When they obtained the spectrum of the optical object, they found unexpected and at first unexplainable emission lines superimposed on a flat continuum. This object remained a mystery until another similar but optically brighter object, 3C 273, was examined in 1963 (Figure 21). Investigators noticed that 3C 273 had a normal spectrum with the same emission lines as observed in radio galaxies, though greatly redshifted (*i.e.,* the spectral lines are displaced to longer wavelengths), as by the Doppler effect. If the redshift were to be ascribed to velocity, however, it would imply an immense velocity of recession. In the case of 3C 48, the redshift had been so large as to shift familiar lines so far that they were not recognized. Many more such objects were found, and they came to be known as quasi-stellar radio sources, abbreviated as quasars.

Large redshifts of quasars

Although the first 20 years of quasar studies were noted more for controversy and mystery than for progress in understanding, the 1980s finally saw a solution to the questions raised by these strange objects. It is now clear that quasars are extreme examples of energetic galaxy nuclei. The amount of radiation emitted by such a nucleus overwhelms the light from the rest of the galaxy so that only very special observational techniques can reveal the galaxy's existence.

A quasar has many remarkable properties. Although it is extremely small (only the size of the solar system), it emits up to 100 times as much radiation as an entire galaxy. It is a complex mixture of very hot gas, cooler gas and dust, and particles that emit synchrotron radiation. Its brightness often varies over short periods—days or even hours. The galaxy underlying the brilliant image of a quasar is

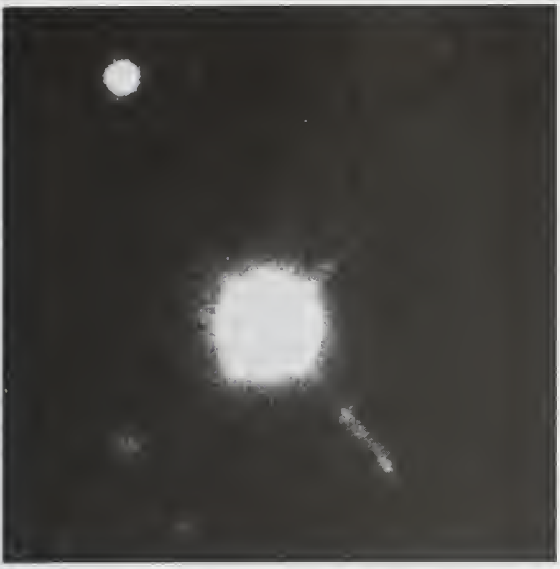


Figure 21: Quasar 3C 273, the brightest and closest of the quasi-stellar radio sources.

By courtesy of the National Optical Astronomy Observatories

probably fairly normal in its properties except for the superficial large-scale effects of the quasar at its centre. Quasars apparently are powered by the same mechanism attributed to radio galaxies. They demonstrate in an extreme way what a supermassive object at the centre of a galaxy can do.

With the gradual recognition of the causes of the quasar phenomenon has come an equally gradual realization that they are simply extreme examples of a process that can be observed in more familiar objects. The black holes that are thought to inhabit the cores of the quasar galaxies are similar to, though more explosive than, those that appear to occur in certain unusual nearer galaxies known as Seyfert Galaxies (see above *Types of galaxies: Principal schemes of classification*). The radio galaxies fall in between. The reason for the differences in the level of activity is apparently related to the source of the gas and stars that are falling into the centres of such objects, providing the black holes with fuel. In the case of quasars, evidence suggests that an encounter with another galaxy, which causes the latter to be tidally destroyed and its matter to fall into the centre of the more massive quasar galaxy, is the cause of its activity. As the material approaches the black hole, it is greatly accelerated, and some of it is expelled by the prevailing high temperatures and drastically rapid motions. This process probably also explains the impressive but lower-level activity in the nuclei of radio and Seyfert galaxies. The captured mass may be of lesser amount—*i.e.*, either a smaller galaxy or a portion of the host galaxy itself. Quasars are more common in that part of the universe observed to have high redshifts, meaning that they were more common about 10 or so billion years ago than they are now, in all probability a result of the higher density of galaxies at that time.

Evolution of galaxies and quasars

The study of the origin and evolution of galaxies and the quasar phenomenon has only just begun. Many models of galaxy formation and evolution have been constructed on the basis of assumptions about conditions in the early universe, which are in turn based on models of the expansion of the Cosmos after the "big bang"—the primordial explosion from which the Cosmos is thought to have originated. Prevailing theory has it that at crucial points in time there condensed from the expanding matter smaller clouds (protogalaxies) that could collapse under their own gravitational field and eventually form galaxies. At the time when the mass of such a stable perturbation in the cloud was approximately 10^{12} solar masses, the galaxies formed. It is still not known whether the clusters of galaxies emerged first or whether they resulted as accumulations of already formed galaxies. Following the separation of mass into individual galaxies, the next step probably depended on the characteristics of the particular clump of matter involved, especially on its mass and angular momentum. The latter quantity was the most likely determinant of the form of the galaxy that eventually evolved. It is thought that a protogalaxy with a large amount of angular momentum tended to form a flat, rapidly rotating system (a spiral galaxy), whereas one with very little angular momentum developed into a more nearly spherical system (an elliptical galaxy).

Significance of angular momentum

Calculations show that a galaxy very gradually becomes dimmer and redder as time progresses and its constituent stars evolve. There is some evidence from very distant galaxies—those whose light was emitted billions of years ago when they were younger—that the effects of this kind of slow evolution can actually be seen.

A more spectacular example of galaxy evolution is provided by quasars. From the statistics of the frequency of different redshifts, which represent different distances and different epochs in the past, it has been determined that the quasar phenomenon occurred most frequently a few billion years after the big bang (the exact amount of time is uncertain because astronomers do not as yet know enough about the geometry or the age of the universe). It appears that conditions did not become suitable for quasar formation until after the galaxies had formed and separated. Today quasars are quite rare, but many galaxies have miniature versions of them in their nuclei in the form of less massive yet remarkably energetic objects.

BIBLIOGRAPHY

The Galaxy: BART J. BOK and PRISCILLA F. BOK, *The Milky Way*, 5th ed. (1981), contains a comprehensive and up-to-date account, without mathematics, of modern galactic research. DIMITRI MIHALAS and JAMES BINNEY, *Galactic Astronomy: Structure and Kinematics*, 2nd ed. (1981), is a thorough, mathematically replete textbook on the Milky Way Galaxy and other galaxies, with emphasis on structure and dynamics. Two collections are HUGO VAN WOERDEN, RONALD J. ALLEN, and W. BUTLER BURTON (eds.), *The Milky Way Galaxy* (1985), symposium articles written at a technical level; and ADRIAAN BLAAUW and MAARTEN SCHMIDT (eds.), *Galactic Structure* (1965), a group of fundamental articles, most of which are still useful sources of basic data, methods, and approaches.

External galaxies: EDWIN HUBBLE, *The Realm of the Nebulae* (1936, reprinted 1982), is a classic account of the early history of extragalactic research written by one of the principal investigators. ALLAN SANDAGE, *The Hubble Atlas of Galaxies* (1961), discusses galaxy classification and includes marvelous full-page illustrations of different types of galaxies. ALLAN SANDAGE, MARY SANDAGE, and JEROME KRISTIAN, *Galaxies and the Universe* (1976, reprinted 1982), contains a comprehensive collection of review articles. PAUL W. HODGE, *Galaxies* (1986), is a nonmathematical introduction. PAUL W. HODGE (comp.), *The Universe of Galaxies* (1984), a collection of *Scientific American* articles, covers most topics of modern galactic research. MICHAEL ROWAN-ROBINSON, *The Cosmological Distance Ladder* (1985), elaborately and comprehensively reviews the distance problem. SIDNEY VAN DEN BERGH and CHRISTOPHER J. PRITCHET (eds.), *The Extragalactic Distance Scale* (1988), collects technical symposium research papers of varying lengths.

(P.W.H.)

Galileo

Galileo Galilei, Italian natural philosopher, astronomer, and mathematician, made fundamental contributions to the sciences of motion and astronomy and to the development of the scientific method. His formulation of circular inertia, the law of falling bodies, and parabolic trajectories marked the beginning of a fundamental change in the study of motion. His insistence that the book of nature was written in the language of mathematics changed natural philosophy from a verbal, qualitative account to a mathematical one in which experimentation became a recognized method for discovering the facts of nature. Finally, his discoveries with the telescope revolutionized astronomy and paved the way for the acceptance of the Copernican heliocentric system—though his advocacy of that system eventually resulted in an Inquisition process against him.

© Scala/Art Resource, New York City

Galileo, oil painting by Justus Sustermans, *c.* 1637; in the Uffizi, Florence.

Early life and career. Galileo was born in Pisa, Tuscany, on February 15, 1564, the oldest son of Vincenzo Galilei, a musician who made important contributions to the theory and practice of music and who may have performed some experiments with Galileo in 1588–89 on the relationship between pitch and the tension of strings. In the early 1570s the family moved to Florence, where the Galilei family had lived for generations. In his middle teens Galileo attended the monastery school at Vallombrosa, near Florence, and then in 1581 matriculated at the University of Pisa, where he was to study medicine. However, he became enamoured of mathematics and decided to make the mathematical subjects and philosophy his profession, against the protests of his father. Galileo then began to prepare himself to teach Aristotelian philosophy and mathematics, and several of his lectures have survived. In 1585 Galileo left the university without having obtained a degree, and for several years he gave private lessons in the mathematical subjects in Florence and Siena. During this period he designed a new form of hydrostatic balance for weighing small quantities and wrote a short treatise, *La bilancetta* ("The Little Balance"), that circulated in manuscript form. He also began his studies on motion, which he pursued steadily for the next two decades.

In 1588 Galileo applied for the chair of mathematics at the University of Bologna but was unsuccessful. His reputation was, however, increasing, and later that year he was asked to deliver two lectures to the Florentine Academy, a prestigious literary group, on the arrangement of the world in Dante's *Inferno*. He also found some ingenious theorems on centres of gravity (again, circulated in manuscript) that brought him recognition among mathematicians and the patronage of Guidobaldo del Monte (1545–1607), a nobleman and author of several important works on mechanics. As a result, he obtained the chair of mathematics at the University of Pisa in 1589. There, according to his first biographer, Vincenzo Viviani (1622–1703), Galileo demonstrated, by dropping bodies of different weights from the top of the famous Leaning Tower, that the speed of fall of a heavy object is not proportional to its weight, as Aristotle had claimed. The manuscript tract *De motu* (*On Motion*), finished during this period, shows that Galileo was abandoning Aristotelian notions about motion and was instead taking an Archimedean approach to the problem. But his attacks on Aristotle made him unpopular with his colleagues, and in 1592 his contract was not renewed. His patrons, however, secured him the chair of mathematics at the University of Padua, where he taught from 1592 until 1610.

The experiments at Pisa

Although Galileo's salary was considerably higher there, his responsibilities as the head of the family (his father had died in 1591) meant that he was chronically pressed for money. His university salary could not cover all his expenses, and he therefore took in well-to-do boarding students whom he tutored privately in such subjects as fortification. He also sold a proportional compass, or sector, of his own devising, made by an artisan whom he employed in his house. Perhaps because of these financial problems, he did not marry, but he did have an arrangement with a Venetian woman, Marina Gamba, who bore him two daughters and a son. In the midst of his busy life he continued his research on motion, and by 1609 he had determined that the distance fallen by a body is proportional to the square of the elapsed time (the law of falling bodies) and that the trajectory of a projectile is a parabola, both conclusions that contradicted Aristotelian physics.

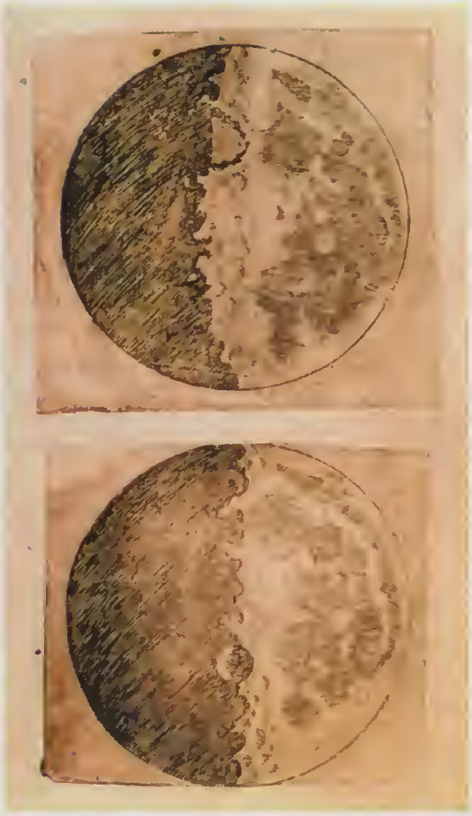
Telescopic discoveries. At this point, however, Galileo's career took a dramatic turn. In the spring of 1609 he heard that in the Netherlands an instrument had been invented that showed distant things as though they were nearby. By trial and error, he quickly figured out the secret of the invention and made his own three-powered spyglass from lenses for sale in spectacle makers' shops. Others had done

© Scala/Art Resource, New York City

Two of Galileo's first telescopes; in the Museum of Science, Florence.

the same; what set Galileo apart was that he quickly figured out how to improve the instrument, taught himself the art of lens grinding, and produced increasingly powerful telescopes. In August of that year he presented an eight-powered instrument to the Venetian Senate (Padua was in the Venetian republic). He was rewarded with life tenure and a doubling of his salary. Galileo was now one of the highest-paid professors at the university. In the fall of 1609 Galileo began observing the heavens with instruments that magnified up to 20 times. In December he drew the Moon's phases as seen through the telescope, showing that the Moon's surface is not smooth, as had been thought, but is rough and uneven. In January 1610 he discovered four moons revolving around Jupiter. He also found that the telescope showed many more stars than are visible with the naked eye. These discoveries were earthshaking, and Galileo quickly produced a little book, *Sidereus Nuncius* (*The Sidereal Messenger*), in which he described them. He dedicated the book to Cosimo II de Medici (1590–1621), the grand duke of his native Tuscany, whom he had tutored in mathematics for several summers, and he named the moons of Jupiter after the Medici family: the Sidera Medicea, or "Medicean Stars." Galileo was rewarded with an appointment as mathematician and philosopher of the grand duke of Tuscany, and in the fall of 1610 he returned in triumph to his native land.

© Scala/Art Resource, New York City



Galileo's illustrations of the Moon, from his *Sidereus Nuncius* (1610; *The Sidereal Messenger*).

The Sun at the centre of the universe

Galileo was now a courtier and lived the life of a gentleman. Before he left Padua he had discovered the puzzling appearance of Saturn, later to be shown as caused by a ring surrounding it, and in Florence he discovered that Venus goes through phases just as the Moon does. Although these discoveries did not prove that the Earth is a planet orbiting the Sun, they undermined Aristotelian cosmology: the absolute difference between the corrupt earthly region and the perfect and unchanging heavens was proved wrong by the mountainous surface of the Moon, the moons of Jupiter showed that there had to be more than one centre of motion in the universe, and the phases of Venus showed that it (and, by implication, Mercury) revolves around the Sun. As a result, Galileo was confirmed in his belief, which he had probably held for decades but which had not been central to his studies, that the Sun is the centre of the universe and that the Earth is a planet, as Copernicus had argued. Galileo's conversion to Copernicanism would be a key turning point in the scientific revolution.

After a brief controversy about floating bodies, Galileo again turned his attention to the heavens and entered a debate with Christoph Scheiner (1573–1650), a German Jesuit and professor of mathematics at Ingolstadt, about the nature of sunspots (of which Galileo was an independent discoverer). This controversy resulted in Galileo's *Istoria e dimostrazioni intorno alle macchie solari e loro accidenti* ("History and Demonstrations Concerning Sunspots and Their Properties," or "Letters on Sunspots"), which appeared in 1613. Against Scheiner, who, in an effort to save the perfection of the Sun, argued that sunspots are satellites of the Sun, Galileo argued that the spots are on or near the Sun's surface, and he bolstered his argument with a series of detailed engravings of his observations.

Galileo's Copernicanism. Galileo's increasingly overt Copernicanism began to cause trouble for him. In 1613 he wrote a letter to his student Benedetto Castelli (1528–1643) in Pisa about the problem of squaring the Copernican theory with certain biblical passages. Inaccurate copies of this letter were sent by Galileo's enemies to the Inquisition in Rome, and he had to retrieve the letter and send an accurate copy. Several Dominican fathers in Florence lodged complaints against Galileo in Rome, and Galileo went to Rome to defend the Copernican cause and his good name. Before leaving, he finished an expanded version of the letter to Castelli, now addressed to the grand duke's mother and good friend of Galileo, the dowager Christina. In his *Letter to the Grand Duchess Christina*, Galileo discussed the problem of interpreting biblical passages with regard to scientific discoveries but, except for one example, did not actually interpret the Bible. That task had been reserved for approved theologians in the wake of the Council of Trent (1545–63) and the beginning of the Catholic Counter-Reformation. But the tide in Rome was turning against the Copernican theory, and in 1615, when the cleric Paolo Antonio Foscarini (*c.* 1565–1616) published a book arguing that the Copernican theory did not conflict with scripture, Inquisition consultants examined the question and pronounced the Copernican theory heretical. Foscarini's book was banned, as were some more technical and nontheological works, such as Johannes Kepler's *Epitome of Copernican Astronomy*. Copernicus's own 1543 book, *De revolutionibus orbium coelestium libri vi* ("Six Books Concerning the Revolutions of the Heavenly Orbs"), was suspended until corrected. Galileo was not mentioned directly in the decree, but he was admonished by Robert Cardinal Bellarmine (1542–1621) not to "hold or defend" the Copernican theory. An improperly prepared document placed in the Inquisition files at this time states that Galileo was admonished "not to hold, teach, or defend" the Copernican theory "in any way whatever, either orally or in writing."

Galileo was thus effectively muzzled on the Copernican issue. Only slowly did he recover from this setback. Through a student, he entered a controversy about the nature of comets occasioned by the appearance of three comets in 1618. After several exchanges, mainly with Orazio Grassi (1583–1654), a professor of mathematics at the Collegio Romano, he finally entered the argument under his own name. *Il saggiatore* (*The Assayer*), published in 1623, was a brilliant polemic on physical reality and an exposition of the new scientific method. Galileo here discussed the method of the newly emerging science, arguing:

The language of nature is mathematics

> Philosophy is written in this grand book, the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures without which it is humanly impossible to understand a single word of it.

He also drew a distinction between the properties of external objects and the sensations they cause in us—*i.e.*, the distinction between primary and secondary qualities. Publication of *Il saggiatore* came at an auspicious moment, for Maffeo Cardinal Barberini (1568–1644), a friend, admirer, and patron of Galileo for a decade, was named Pope Urban VIII as the book was going to press. Galileo's friends quickly arranged to have it dedicated to the new pope. In 1624 Galileo went to Rome and had six interviews with Urban VIII. Galileo told the pope about his theory of the tides (developed earlier), which he put forward as proof of the annual and diurnal motions of the Earth. The pope gave Galileo permission to write a book about theories of the universe but warned him to treat the Copernican theory only hypothetically. The book, *Dialogo sopra i due massimi sistemi del mondo, tolemaico e copernicano* (*Dialogue Concerning the Two Chief World Systems, Ptolemaic & Copernican*), was finished in 1630, and Galileo sent it to the Roman censor. Because of an outbreak of the plague, communications between Florence and Rome were interrupted, and Galileo asked for the censoring to be done instead in Florence. The Roman censor had a number of serious criticisms of the book and forwarded these to his colleagues in Florence. After writing a preface in which he professed that what followed was written hypothetically, Galileo had little trouble getting the book through the Florentine censors, and it appeared in Florence in 1632.

Galileo before the Inquisition. In the *Dialogue*'s witty conversation between Salviati (representing Galileo), Sagredo (the intelligent layman), and Simplicio (the dyed-in-the-wool Aristotelian), Galileo gathered together all the arguments (mostly based on his own telescopic discoveries) for the Copernican theory and against the traditional geocentric cosmology. As opposed to Aristotle's, Galileo's approach to cosmology is fundamentally spatial and geometric: the Earth's axis retains its orientation in space as the Earth circles the Sun, and bodies not under a force retain their velocity (although this inertia is ultimately circular). But in giving Simplicio the final word, that God could have made the universe any way he wanted to and still made it appear to us the way it does, he put Pope Urban VIII's favourite argument in the mouth of the person who had been ridiculed throughout the dialogue. The reaction against the book was swift. The pope convened a special commission to examine the book and make recommendations; the commission found that Galileo had not really treated the Copernican theory hypothetically and recommended that a case be brought against him by the Inquisition. Galileo was summoned to Rome in 1633. During his first appearance before the Inquisition, he was confronted with the 1616 edict recording that he was forbidden to discuss the Copernican theory. In his defense Galileo produced a letter from Cardinal Bellarmine, by then dead, stating that he was admonished only not to hold or defend the theory. The case was at somewhat of an impasse, and, in what can only be called a plea bargain, Galileo confessed to having overstated his case. He was pronounced to be vehemently suspect of heresy and was condemned to life imprisonment and was made to abjure formally. There is no evidence that at this time he whispered, "Eppur si muove" ("And yet it moves"). It should be noted that Galileo was never in a dungeon or tortured; during the Inquisition process he stayed mostly at the house of the Tuscan ambassador to the Vatican and for a short time in a comfortable apartment in the Inquisition building. After the process he spent six months at the palace of Ascanio Piccolomini (*c.* 1590–1671), the archbishop of Siena and a friend and patron, and then moved into a villa near Arcetri, in the hills above Florence. He spent the rest of his life there. Galileo's daughter Sister Maria Celeste, who was in a nearby nunnery, was a great comfort to her father until her untimely death in 1634.

Under house arrest

Galileo was then 70 years old. Yet he kept working. In Siena he had begun a new book on the sciences of motion and strength of materials. There he wrote up his unpublished studies that had been interrupted by his interest in the telescope in 1609 and pursued intermittently since. The book was spirited out of Italy and published in Leiden, Netherlands, in 1638 under the title *Discorsi e dimostrazioni matematiche intorno a due nuove scienze attenenti alla meccanica* (*Dialogues Concerning Two New Sciences*). Galileo here treated for the first time the bending and breaking of beams and summarized his mathematical and experimental investigations of motion, including the law of falling bodies and the parabolic path of projectiles as a result of the mixing of two motions, constant speed and uniform acceleration. By then Galileo had become blind, and he spent his time working with a young student, Vincenzo Viviani, who was with him when he died at Arcetri on January 8, 1642.

BIBLIOGRAPHY. The standard edition of Galileo's works is *Le opere di Galileo Galilei*, 20 vol. in 21, ed. by ANTONIO FAVARO (1890–1909, reissued 1968). There are now several English translations of *Sidereus Nuncius* (1610): *The Sidereal Messenger of Galileo Galilei and a Part of the Preface to Kepler's Dioptrics Containing the Original Account of Galileo's Astronomical Discoveries*, trans. by EDWARD STAFFORD CARLOS (1880, reprinted 1959); *Sidereus Nuncius; or, The Sidereal Messenger*, trans. by ALBERT VAN HELDEN (1989); and STILLMAN DRAKE, *Telescopes, Tides, and Tactics: A Galilean Dialogue About the Starry Messenger and Systems of the World* (1983), which contains a complete translation interspersed in the dialogue. *Discorso…intorno alle cose, che stanno iǹ sù l'acqua, ò che in quella si muovono* (1612) appears in English as *Discourse on Bodies in Water*, trans. by THOMAS SALUSBURY and ed. by STILLMAN DRAKE (1960); and is interspersed in a dialogue in STILLMAN DRAKE, *Cause, Experiment, and Science: A Galilean Dialogue, Incorporating a New English Translation of Galileo's Bodies That Stay Atop Water, or Move in It* (1981). Galileo's letter to the grand duchess Christina on the relationship of science to religion, written in 1615, was published in Latin in Strasbourg in 1636 without Galileo's permission; English translations are "Letter to Madame Christina…," in *Discoveries and Opinions of Galileo*, trans. by STILLMAN DRAKE (1957, reissued 1990), pp. 175–216, which also includes abridged translations of *Sidereus Nuncius*, the letters on sunspots (*Istoria e dimostrazioni intorno alle macchie solari e loro accidenti*, 1613), and *Il saggiatore;* and "Galileo's Letter to the Grand Duchess Christina (1615)," in MAURICE A. FINOCCHIARO (ed. and trans.), *The Galileo Affair: A Documentary History* (1989), pp. 87–118, which also includes the translated documents of Galileo's trial. The standard translation of *Dialogo sopra i due massimi sistemi del mondo tolemaico, e copernicano* (1632) is *Dialogue Concerning the Two Chief World Systems, Ptolemaic & Copernican*, trans. by STILLMAN DRAKE, 2nd ed. (1967). Galileo's *Discorsi e dimostrazioni matematiche, intorno à due nuove scienze . . .* (1638) appears as *Dialogues Concerning Two New Sciences*, trans. by HENRY CREW and ALFONSO DE SALVIO (1914, reissued 1968); and *Two New Sciences*, trans. by STILLMAN DRAKE, rev. ed. (1989). The correspondence between Galileo and one of his daughters is available in MARY ALLAN-OLNEY (compiler), *The Private Life of Galileo: Compiled Principally from his Correspondence and That of His Eldest Daughter, Sister Maria Celeste* (1870).

Several biographies of Galileo have been written by STILLMAN DRAKE: *Galileo at Work: His Scientific Biography* (1978, reissued 1995), *Galileo: Pioneer Scientist* (1990), and *Galileo* (1980). JAMES RESTON, JR., *Galileo* (1994), is a well-documented popular biography. Portraits and other depictions of Galileo can be found in J.J. FAHIE, *Memorials of Galileo Galilei, 1564–1642* (1929).

Studies of various aspects of Galileo's life and career include MAURICE A. FINOCCHIARO, *Galileo and the Art of Reasoning: Rhetorical Foundations of Logic and Scientific Method* (1980); SILVIO A. BEDINI, *The Pulse of Time: Galileo Galilei, the Determination of Longitude, and the Pendulum Clock* (1991); MICHAEL SEGRE, *In the Wake of Galileo* (1991); VICTOR COELHO (ed.), *Music and Science in the Age of Galileo* (1992); JOSEPH C. PITT, *Galileo, Human Knowledge, and the Book of Nature: Method Replaces Metaphysics* (1992); MARIO BIAGIOLI, *Galileo, Courtier: The Practice of Science in the Culture of Absolutism* (1993); JEAN DIETZ MOSS, *Novelties in the Heavens: Rhetoric and Science in the Copernican Controversy* (1993); and Cesare S. Maffioli, *Out of Galileo: The Science of Waters, 1628–1718* (1994).

Works specifically treating Galileo and the Roman Catholic church include KARL VON GEBLER, *Galileo Galilei and the Roman Curia* (1879, reprinted 1977; originally published in German, 2 vol., 1876–77); GIORGIO DE SANTILLANA, *The Crime of Galileo* (1955, reprinted 1981); JEROME J. LANGFORD, *Galileo, Science, and the Church*, 3rd ed. (1992); PIETRO REDONDI, *Galileo Heretic* (1987; originally published in Italian, 1983); RICHARD S. WESTFALL, *Essays on the Trial of Galileo* (1989); RICHARD J. BLACKWELL, *Galileo, Bellarmine, and the Bible* (1991); RIVKA FELDHAY, *Galileo and the Church: Political Inquisition or Critical Dialogue?* (1995); and ANNIBALE FANTOLI, *Galileo: For Copernicanism and for the Church*, 2nd ed., rev. and corrected (1996; originally published in Italian, 1993).

(A.V.H.)

Game Theory

Game theory is a branch of applied mathematics that provides tools for analyzing situations in which parties, called players, make decisions that are interdependent. This interdependence causes each player to consider the other player's possible decisions, or strategies, in formulating his own strategy. A solution to a game describes the optimal decisions of the players, who may have similar, opposed, or mixed interests, and the outcomes that may result from these decisions.

Although game theory can be and has been used to analyze parlour games, its applications are much broader. In fact, game theory was originally developed by the Hungarian-born American mathematician John von Neumann and his Princeton University colleague Oskar Morgenstern, a German-born American economist, to solve problems in economics. In their book *Theory of Games and Economic Behavior* (1944), von Neumann and Morgenstern asserted that the mathematics developed for the physical sciences, which describes the workings of a disinterested nature, was a poor model for economics. They observed that economics is much like a game, wherein players anticipate each other's moves, and therefore requires a new kind of mathematics, which they called game theory.

Game theory has been applied to a wide variety of situations in which the choices of players interact to affect the outcome. In stressing the strategic aspects of decision making, the theory both supplements and goes beyond the classical theory of probability. It has been used, for example, to determine what political coalitions or business conglomerates are likely to form, the optimal price at which to sell products or services in the face of competition, the best site for a manufacturing plant, and the behaviour of certain life-forms in their struggle for survival. It has even been used to challenge the legality of certain voting systems.

It would be surprising if any one theory could address such an enormous range of "games," and in fact there is no single game theory. A number of theories have been proposed, each applicable to different situations and each with its own concepts of what constitutes a solution. This article describes some simple games, discusses different theories, and outlines principles underlying game theory.

The article is divided into the following sections:

CLASSIFICATION OF GAMES

Games can be classified according to certain significant features, the most obvious of which is the number of players. Thus, a game can be designated as being a one-person, two-person, or *n*-person (with *n* greater than two) game, with games in each category having their own distinctive features. In addition, a player need not be an individual; it may be a nation, a corporation, or a team comprising many people with shared interests.

In games of perfect information, such as chess, each player knows everything about the game at all times. Poker, on the other hand, is a game of imperfect information because players do not know all of their opponents' cards.

Constant-sum games

The extent to which the goals of the players coincide or conflict is another basis for classifying games. Constant-sum games are games of total conflict, which are also called games of pure competition. Poker, for example, is a constant-sum game because the combined wealth of the players remains constant.

Players in constant-sum games have completely opposed interests, whereas in variable-sum games they may all be winners or losers. In a labour-management dispute, for example, the two parties certainly have some conflicting interests, but both will benefit if a strike is averted.

Variable-sum games

Variable-sum games can be further distinguished as being either cooperative or noncooperative. In cooperative games players can communicate and make binding agreements; in noncooperative games players may communicate, but they cannot make binding agreements, such as an enforceable contract. An automobile salesperson and a potential customer will be engaged in a cooperative game if they agree on a price and sign a contract. However, the dickering that they do to reach this point will be noncooperative. Similarly, when people bid independently at an auction they are playing a noncooperative game, even though the high bidder agrees to complete the purchase.

Finally, a game is said to be finite when each player has a finite number of options, the number of players is finite, and the game cannot go on indefinitely. Chess, checkers, poker, and most parlour games are finite. Infinite games are more subtle and will only be touched upon in this article.

A game can be described in extensive, normal, or characteristic-function form. Most parlour games, which progress one move at a time, can be modeled as games in extensive form. Extensive-form games can be described by a "game tree," in which each turn is a vertex of the tree, with each branch indicating the players' successive choices.

The normal (strategic) form is primarily used to describe two-person games. In this form a game is represented by a payoff matrix, wherein each row describes the strategy of one player and each column describes the strategy of the other player. The matrix entry at the intersection of each row and column gives the outcome of each player choosing the corresponding strategy. The payoffs to each player associated with this outcome are the basis for determining whether the strategies are "in equilibrium," or stable.

The characteristic-function form is generally used to analyze games with more than two players. It indicates the minimum value that each coalition of players—including single-player coalitions—can guarantee for itself when playing against a coalition made up of all the other players.

ONE-PERSON GAMES

With no opponents, the player only needs to list available options and then choose the optimal outcome in one-person games. When chance is involved the game might seem to be more complicated, but in principle the decision is still relatively simple. For example, a person deciding whether to carry an umbrella weighs the costs and benefits of carrying or not carrying it. While this person may make the wrong decision, there does not exist a conscious opponent. That is, nature is presumed to be completely indifferent to the player's decision, and the person can base his decision on simple probabilities. One-person games hold little interest for game theorists.

TWO-PERSON CONSTANT-SUM GAMES

Strictly determined

Games of perfect information. The simplest game of any real theoretical interest is a two-person constant-sum game of perfect information. Examples of such games include chess, checkers, and the Japanese game of go. In 1912 the German mathematician Ernst Zermelo proved that such games are strictly determined; by making use of all available information, the players can deduce strategies that are optimal, which makes the outcome preordained. In chess, for example, exactly one of three outcomes must occur if the players make optimal choices: (1) White wins (has a strategy that wins against any strategy of Black); (2) Black wins; or (3) White and Black draw. In principle, a sufficiently powerful computer could determine which of the three outcomes will occur. However, considering that there are some 10^{43} distinct 40-move games of chess possible, there seems no possibility that such a computer will be developed in the foreseeable future.

Games of imperfect information. A "saddlepoint" in a two-person constant-sum game is the outcome that rational players would choose. (Its name derives from its being the minimum of a row that is also the maximum of a column in a payoff matrix—to be illustrated shortly—which corresponds to the shape of a saddle.) A saddlepoint always exists in games of perfect information but may not exist in games of imperfect information. By choosing a strategy associated with this outcome, each player obtains an amount at least equal to his payoff at that outcome, no matter what the other player does. This payoff is called the value of the game; as in perfect-information games, it is preordained by the players' choices of strategies associated with the saddlepoint, making such games strictly determined.

Value of the game

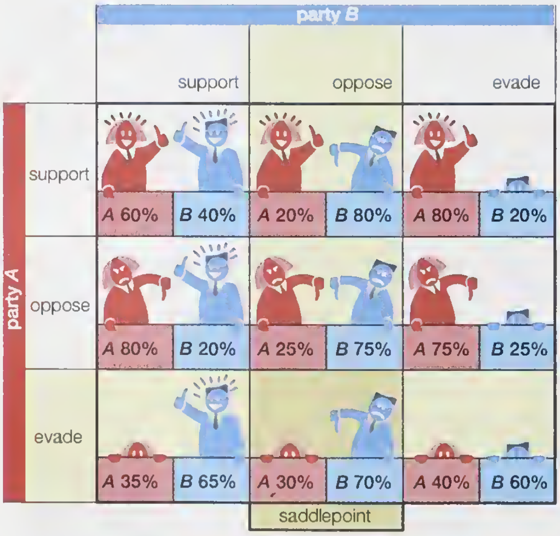
The normal-form game in Table 1 is used to illustrate the calculation of a saddlepoint. Two political parties, A and B, must each decide how to handle a controversial issue in a certain election. Each party can either support the issue, oppose it, or evade it by being ambiguous. The decisions by A and B on this issue determine the percentage of the vote that each party receives. The entries in the payoff matrix represent party A's percentage of the vote (the remaining percentage goes to B). When, for example, A supports the issue and B evades it, A gets 80 percent and B 20 percent of the vote.

Assume that each party wants to maximize its vote. A's decision seems difficult at first because it depends on B's choice of strategy. A does best to support if B evades, oppose if B supports, and evade if B opposes. A must therefore consider B's decision before making its own. Note that no matter what A does, B obtains the largest percentage of the vote (smallest percentage for A) by opposing the issue rather than supporting it or evading it. Once A recognizes this, its strategy obviously should be to evade, settling for 30 percent of the vote. Thus, a 30 to 70 percent division of the vote, to A and B respectively, is the game's saddlepoint.

Maximin and minimax value

A more systematic way of finding a saddlepoint is to determine the so-called maximin and minimax values. A first determines the minimum percentage of votes it can obtain for each of its strategies; it then finds the maximum of these three minimum values, giving the maximin. The minimum percentages A will get if it supports, opposes, or evades are, respectively, 20, 25, and 30. The largest of these, 30, is the maximin value. Similarly, for each strategy B chooses, it determines the maximum percentage of votes A will win (and thus the minimum that it can win). In this case, if B supports, opposes, or evades, the maximum A will get is 80, 30, and 80, respectively. B will obtain its largest percentage by minimizing A's maximum percent of the vote, giving the minimax. The smallest of A's maximum values is 30, so 30 is B's minimax value. Because both the minimax and the maximin values coincide, 30 is a saddlepoint. The two parties might as well announce their strategies in advance, because the other party cannot gain from this knowledge.

Payoff matrix with saddlepoint

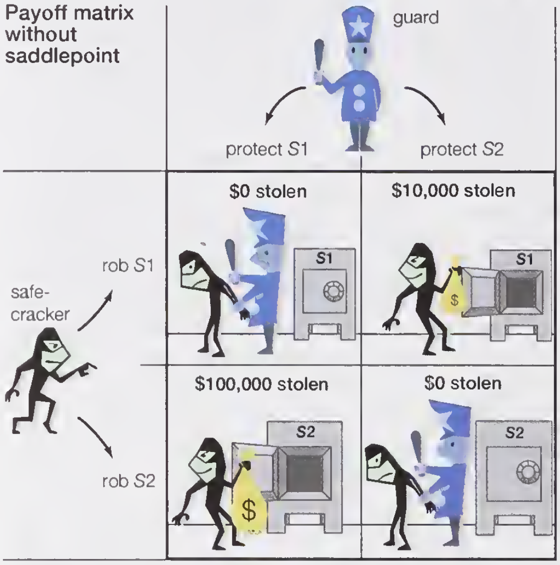


		party B		
		support	oppose	evade
party A	support	A 60% B 40%	A 20% B 80%	A 80% B 20%
	oppose	A 80% B 20%	A 25% B 75%	A 75% B 25%
	evade	A 35% B 65%	A 30% B 70% saddlepoint	A 40% B 60%

Table 1.

Mixed strategies and the minimax theorem. When saddlepoints exist, the optimal strategies and outcomes can be easily determined. However, when there is no saddlepoint the calculation is more elaborate, as illustrated in Table 2.

Payoff matrix without saddlepoint



		guard	
		protect $S1$	protect $S2$
safecracker	rob $S1$	$0 stolen	$10,000 stolen
	rob $S2$	$100,000 stolen	$0 stolen

Table 2.

A guard is hired to protect two safes in separate locations: $S1$ contains \$10,000 and $S2$ contains \$100,000. The guard can protect only one safe at a time. A safecracker and the guard must decide in advance, without knowing what the other party will do, which safe to try to rob and which safe to protect. When they go to the same safe, the safecracker gets nothing; when they go to different safes, the safecracker gets the contents of the unprotected safe.

In such a game, game theory does not indicate that any one particular strategy is best. Instead, it prescribes that a strategy be chosen in accordance with a probability distribution, which in this simple example is quite easy to calculate. In larger and more complex games, finding this strategy involves solving a problem in linear programming, which can be considerably more difficult.

To calculate the appropriate probability distribution in this example, each player adopts a strategy that makes him indifferent to what his opponent does. Assume that the guard protects $S1$ with probability p and $S2$ with probability $1 - p$. Thus, if the safecracker tries $S1$, he will be successful whenever the guard protects $S2$. In other words, he will get \$10,000 with probability $1 - p$ and \$0 with probability p for an average gain of $\$10,000(1 - p)$. Similarly, if the safecracker tries $S2$, he will get \$100,000 with probability p and \$0 with probability $1 - p$ for an average gain of $\$100,000p$.

The guard will be indifferent to which safe the safecracker chooses if the average amount stolen is the same in both cases—that is, if $\$10,000(1 - p) = \$100,000p$. Solving for p gives $p = 1/11$. If the guard protects $S1$ with probability $1/11$ and $S2$ with probability $10/11$, he will lose, on average, no more than about \$9,091 whatever the safecracker does.

Using the same kind of argument, it can be shown that the safecracker will get an average of at least \$9,091 if he tries to steal from $S1$ with probability $10/11$ and from $S2$ with probability $1/11$. This solution in terms of mixed strategies, which are assumed to be chosen at random with the indicated probabilities, is analogous to the solution of the game with a saddlepoint (in which a pure, or single best, strategy exists for each player).

Transparent strategies

The safecracker and the guard give away nothing if they announce the probabilities with which they will randomly choose their respective strategies. On the other hand, if they make themselves predictable by exhibiting any kind of pattern in their choices, this information can be exploited by the other player.

The minimax theorem, which von Neumann proved in 1928, states that every finite, two-person constant-sum game has a solution in pure or mixed strategies. Specifically, it says that for every such game between players A and B, there is a value v and strategies for A and B such that, if A adopts its optimal (maximin) strategy, the outcome will be at least as favourable to A as v; if B adopts its optimal (minimax) strategy, the outcome will be no more favourable to A than v. Thus, A and B have both the incentive and the ability to enforce an outcome that gives an (expected) payoff of v.

Utility theory. In the previous example it was tacitly assumed that the players were maximizing their average profits, but in practice players may consider other factors. For example, few people would risk a sure gain of $1,000,000 for an even chance of winning either $3,000,000 or $0, even though the expected (average) gain from this bet is $1,500,000. In fact, many decisions that people make, such as buying insurance policies, playing lotteries, and gambling at a casino, indicate that they are not maximizing their average profits. Game theory does not attempt to state what a player's goal should be; instead, it shows how a player can best achieve his goal, whatever that goal is.

Rational axioms

Von Neumann and Morgenstern understood this distinction; to accommodate all players, whatever their goals, they constructed a theory of utility. They began by listing certain axioms that they thought all rational decision makers would follow (for example, if a person likes tea better than coffee, and coffee better than milk, then that person should like tea better than milk). They then proved that it was possible to define a utility function for such decision makers that would reflect their preferences. In essence, a utility function assigns a number to each player's alternatives to convey their relative attractiveness. Maximizing someone's expected utility automatically determines a player's most preferred option. In recent years, however, some doubt has been raised about whether people actually behave in accordance with these axioms, and alternative axioms have been proposed.

TWO-PERSON VARIABLE-SUM GAMES

Much of the early work in game theory was on two-person constant-sum games because they are the easiest to treat mathematically. The players in such games have diametrically opposed interests, and there is a consensus about what constitutes a solution (as given by the minimax theorem). Most games that arise in practice, however, are variable-sum games; the players have both common and opposed interests. For example, a buyer and a seller are engaged in a variable-sum game (the buyer wants a low price and the seller a high one, but both want to make a deal), as are two hostile nations (they may disagree about numerous issues, but both gain if they avoid going to war).

Some "obvious" properties of two-person constant-sum games are not valid in variable-sum games. In constant-sum games both players cannot gain (they may or may not lose, but they cannot both gain) if they are deprived of some of their strategies. In variable-sum games, however, players may gain if some of their strategies are no longer available. This might not seem possible at first. One would think that if a player benefited from not using certain strategies, the player would simply avoid those strategies and choose more advantageous ones, but this is not always the case. For example, in a region with high unemployment a worker may be willing to accept a lower salary to obtain or keep a job, but if a minimum wage law makes that option illegal, the worker may be "forced" to accept a higher salary.

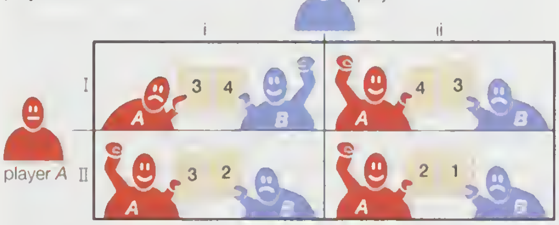
Effect of open communications

The effect of communication is particularly revealing of the difference between constant-sum and variable-sum games. In constant-sum games it never helps a player to give an adversary information, and it never hurts a player to learn an opponent's optimal strategy (pure or mixed) in advance. However, these properties do not necessarily hold in variable-sum games. Indeed, a player may want an opponent to be well-informed. In a labour-management dispute, for example, if the labour union is prepared to strike, it behooves the union to inform management and thereby possibly achieve its goal without a strike. In this example, management is not harmed by the advance information (it, too, benefits by avoiding a costly strike). In other variable-sum games, knowing an opponent's strategy can sometimes be disadvantageous. For example, a blackmailer can only benefit if he first informs his victim that he will harm him—generally by disclosing some sensitive and secret details of the victim's life—if his terms are not met. For such a threat to be credible, the victim must fear the disclosure and believe that the blackmailer is capable of executing the threat. (The credibility of threats is a question that game theory studies.) Although a blackmailer may be able to harm a victim without any communication taking place, a blackmailer cannot extort a victim unless he first adequately informs the victim of his intent and its consequences. Thus, the victim's knowledge of the blackmailer's strategy, including his ability and will to carry out the threat, works to the blackmailer's advantage.

Cooperative versus noncooperative games. Communication is pointless in constant-sum games because there is no possibility of mutual gain from cooperating. In variable-sum games, on the other hand, the ability to communicate, the degree of communication, and even the order in which players communicate can have a profound influence on the outcome.

In the variable-sum game shown in Table 3, each matrix entry consists of two numbers. (Because the combined wealth of the players is not constant, it is impossible to deduce one player's payoff from the payoff of the other; consequently, both players' payoffs must be given.) The first number in each entry is the payoff to the row player (player A), and the second number is the payoff to the column player (player B).

Variable-sum payoff matrix


player B
player A
A
B

player A \ player B	i	ii
I	3 4	4 3
II	3 2	2 1

Table 3.

In this example it will be to player B's advantage if the game is cooperative and to player A's advantage if the game is noncooperative. Without communication, assume each player applies the "sure-thing" principle: it maximizes its minimum payoff by determining the minimum it will receive whatever its opponent does. Thereby, A determines that it will do best to choose strategy I no matter what B does: if B chooses i, A will get 3 regardless of what A does; if B chooses ii, A will get 4 rather than 3. B similarly determines that it will do best to choose i no matter what A does. Selecting these two strategies, A will get 3 and B will get 4 at (3, 4).

Credible threats

In a cooperative game, however, A can threaten to play II unless B agrees to play ii. If B agrees, its payoff will be reduced to 3 while A's payoff will rise to 4 at (4, 3); if B does not agree and A carries out its threat, A will neither gain nor lose at (3, 2) compared to (3, 4), but B will get a payoff of only 2. Clearly, A will be unaffected if B does not agree and thus has a credible threat; B will be affected and obviously will do better at (4, 3) than at (3, 2) and should comply with the threat.

Sometimes both players can gain from the ability to communicate. Two pilots trying to avoid a midair collision clearly will benefit if they can communicate, and the degree of communication allowed between them may even determine whether or not they will crash. Generally, the more two players' interests coincide, the more important and advantageous communication becomes.

The solution to a cooperative game in which players have a common goal involves coordinating the players' decisions effectively. This is relatively straightforward, as is finding the solution to constant-sum games with a saddle-point. For games in which the players have both common and conflicting interests—in other words, in most variable-sum games, whether cooperative or noncooperative—what constitutes a solution is much harder to define and make persuasive.

The Nash solution. Sometimes solutions to variable-sum games seem inequitable or are not enforceable. One well-known cooperative solution to two-person variable-sum games was proposed by the American mathematician John F. Nash, who received the Nobel Prize for Economics in 1994 for this and related work he did in game theory.

Pareto-optimality

Given a game with a set of possible outcomes and associated utilities for each player, Nash showed that there is a unique outcome that satisfies four conditions: (1) The outcome is independent of the choice of a utility function (that is, if a player prefers x to y, the solution will not change if one function assigns x a utility of 10 and y a utility of 1 or a second function assigns the values of 20 and 2). (2) Both players cannot do better simultaneously (a condition known as Pareto-optimality). (3) The outcome is independent of irrelevant alternatives (in other words, if unattractive options are added to or dropped from the list of alternatives, the solution will not change). (4) The outcome is symmetrical (that is, if the players reverse their roles, the solution will remain the same, except that the payoffs will be reversed).

In some cases the Nash solution seems inequitable because it is based on a balance of threats—the possibility that no agreement will be reached, so that both players will suffer losses—rather than a "fair" outcome. When, for example, a rich person and a poor person are to receive $10,000 provided they can agree on how to divide the money (if they fail to agree, they receive nothing), most people assume that the fair solution would be for each person to get half, or even that the poor person should get more than half. According to the Nash solution, however, there is a utility for each player associated with all possible outcomes. Moreover, the specific choice of utility functions should not affect the solution (condition 1) as long as they reflect each person's preferences. In this example, assume that the rich person's utility is equal to one-half the money received and that the poor person's utility is equal to the money received. These different functions reflect the fact that additional income is more precious to the poor person. Under the Nash solution, the threat of reaching no agreement induces the poor person to accept one-third of the $10,000, giving the rich person two-thirds. In general, the Nash solution finds an outcome such that each player gains the same amount of utility.

The Prisoners' Dilemma. To illustrate the kinds of difficulties that arise in two-person noncooperative variable-sum games, consider the celebrated Prisoners' Dilemma (PD), originally formulated by the American mathematician Albert W. Tucker. Two prisoners, A and B, suspected of committing a robbery together, are isolated and urged to confess. Each is concerned only with getting the shortest possible prison sentence for himself; each must decide whether to confess without knowing his partner's decision. Both prisoners, however, know the consequences of their decisions: (1) if both confess, both go to jail for five years; (2) if neither confesses, both go to jail for one year (for carrying concealed weapons); and (3) if one confesses while the other does not, the confessor goes free (for turning state's evidence) and the silent one goes to jail for 20 years. The normal form of this game is shown in Table 4.

Code of silence

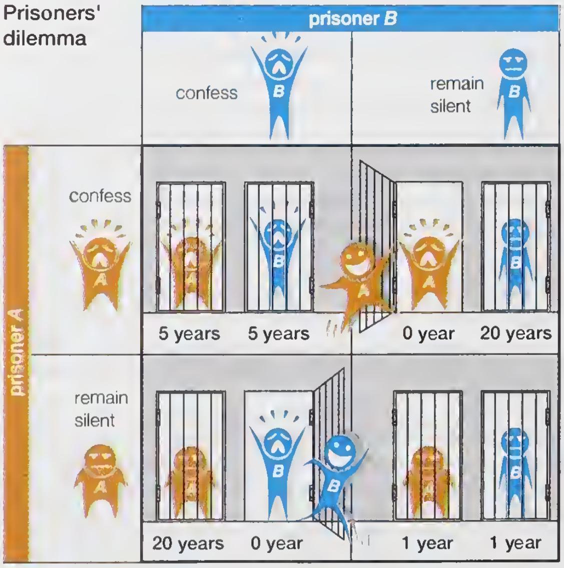
Superficially, the analysis of PD is very simple. Although A cannot be sure what B will do, he knows that he does best to confess when B confesses (he gets five years rather than 20) and also when B remains silent (he serves no time rather than a year); analogously, B will reach the same conclusion. So the solution would seem to be that each prisoner does best to confess and go to jail for five years. Paradoxically, however, the two robbers would do better if they both adopted the apparently irrational strategy of remaining silent; each would then serve only one year in jail. The irony of PD is that when each of two (or more) parties acts selfishly and does not cooperate with the other (that is, when he confesses), they do worse than when they act unselfishly and cooperate together (remain silent).

PD is not just an intriguing hypothetical problem; similar real-life situations have often been observed. For example, two shopkeepers engaged in a price war may well be caught up in a PD. Each shopkeeper knows that if he has lower prices than his rival, he will attract his rival's customers and thereby increase his own profits. Each therefore decides to lower his prices, with the result that neither gains any customers and both earn smaller profits. Similarly, nations competing in an arms race and farmers increasing crop production can also be seen as manifestations of PD. When two nations keep buying more weapons in an attempt to achieve military superiority, neither gains an advantage and both are poorer than when they started. A single farmer can increase his profits by increasing production, but when all farmers increase their output a market glut ensues, with lower profits for all.

It might seem that the paradox inherent in PD could be resolved if the game were played repeatedly. Players would learn that they do best when both act unselfishly and cooperate. Indeed, if one player failed to cooperate in one game, the other player could retaliate by not cooperating in the next game, and both would lose until they began to "see the light" and cooperated again. When the game is repeated a fixed number of times, however, this argument fails. To see this, suppose two shopkeepers set up their booths at a 10-day county fair. Furthermore, suppose that each maintains full prices, knowing that if he does not, his competitor will retaliate the next day. On the last day, however, each shopkeeper realizes that his competitor can no longer retaliate and so there is little reason not to lower their prices. But if each shopkeeper knows that his rival will lower his prices on the last day, he has no incentive to maintain full prices on the ninth day. Continuing this reasoning, one concludes that rational shopkeepers will have a price war every day. It is only when the game is played repeatedly, and neither player knows when the sequence will end, that the cooperative strategy can succeed.

"Nice" strategies

In 1980 the American political scientist Robert Axelrod engaged a number of game theorists in a round-robin tournament. In each match the strategies of two theorists, incorporated in computer programs, competed against one another in a sequence of PDs with no definite end. A "nice" strategy was defined as one in which a player always cooperates with a cooperative opponent. Also, if a player's opponent did not cooperate during one turn, most strategies prescribed noncooperation on the next turn, but a player with a "forgiving" strategy reverted rapidly to cooperation once its opponent started cooperating again. In this experiment it turned out that every nice strategy outperformed every strategy that was not nice. Furthermore, of the nice strategies, the forgiving ones performed best.



		prisoner B	
		confess	remain silent
prisoner A	confess	5 years 5 years	0 year 20 years
	remain silent	20 years 0 year	1 year 1 year

Table 4.

Theory of moves. Another approach to inducing cooperation in variable-sum games is the theory of moves (TOM). Proposed by the American political scientist Steven J. Brams, TOM allows players, starting at any outcome in a payoff matrix, to move and countermove within the matrix, thereby capturing the changing strategic nature of games as they evolve over time. In particular, TOM assumes that players think ahead about the consequences of all of the participants' moves and countermoves when formulating plans. Thereby, TOM embeds extensive-form calculations within the normal form, deriving advantages of both forms: the nonmyopic thinking of the extensive form disciplined by the economy of the normal form.

To illustrate the nonmyopic perspective of TOM, consider what happens in PD as a function of where play starts:

When play starts noncooperatively, players are stuck, no matter how far ahead they look, because as soon as one player departs, the other player, enjoying his best outcome, will not move on. Outcome: The players stay at the noncooperative outcome.

When play starts cooperatively, neither player will defect, because if he does, the other player will also defect, and they both will end up worse off. Thinking ahead, therefore, neither player will defect. Outcome: The players stay at the cooperative outcome.

When play starts at one of the win-lose outcomes (best for one player, worst for the other), the player doing best will know that if he is not magnanimous, and consequently does not move to the cooperative outcome, his opponent will move to the noncooperative outcome, inflicting on the best-off player his next-worst outcome. Therefore, it is in the best-off player's interest, as well as his opponent's, that he act magnanimously, anticipating that if he does not, the noncooperative outcome (next-worst for both), rather than the cooperative outcome (next-best for both), will be chosen. Outcome: The best-off player will move to the cooperative outcome, where play will remain.

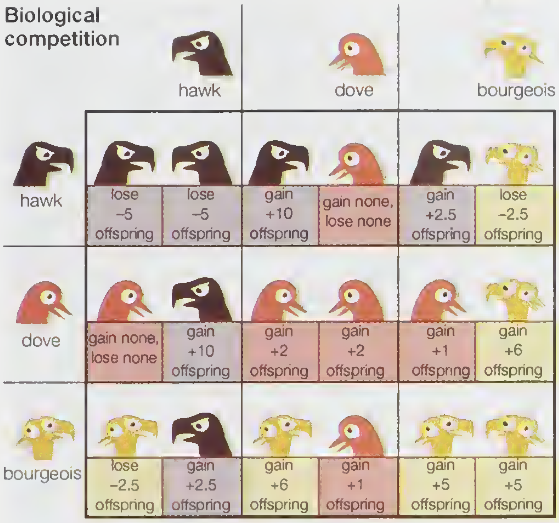
Rational moves

Such rational moves are not beyond the pale of most players. Indeed, they are frequently made by those who look beyond the immediate consequences of their own choices. Such far-sighted players can escape the dilemma in PD—and in other variable-sum games—provided play does not begin noncooperatively. Hence, TOM does not predict unconditional cooperation in PD but, instead, makes it a function of the starting point of play.

Biological applications. One fascinating and unexpected application of game theory in general, and PD in particular, occurs in biology. When two males confront each other, whether competing for a mate or for some disputed territory, they can behave either like "hawks"—fighting until one is maimed, killed, or flees—or like "doves"—posturing a bit but leaving before any serious harm is done. (In effect, the doves cooperate while the hawks do not.) Neither type of behaviour is ideal for survival: a species containing only hawks would have a high casualty rate; a species containing only doves would be vulnerable to an invasion by hawks or a mutation that produces hawks, because the population growth rate of the competitive hawks would be much higher initially than that of the doves.

"Bourgeois" behaviour

Thus, a species with males consisting exclusively of either hawks or doves is vulnerable. The English biologist John Maynard Smith showed that a third type of male behaviour, which he called "bourgeois," would be more stable than that of either pure hawks or pure doves. A bourgeois may act like either a hawk or a dove, depending on some external cues; for example, it may fight tenaciously when it meets a rival in its own territory but yield when it meets the same rival elsewhere. In effect, bourgeois animals submit their conflict to external arbitration to avoid a prolonged and mutually destructive struggle.



Biological competition

	hawk		dove		bourgeois	
hawk	lose −5 offspring	lose −5 offspring	gain +10 offspring	gain none, lose none	gain +2.5 offspring	lose −2.5 offspring
dove	gain none, lose none	gain +10 offspring	gain +2 offspring	gain +2 offspring	gain +1 offspring	gain +6 offspring
bourgeois	lose −2.5 offspring	gain +2.5 offspring	gain +6 offspring	gain +1 offspring	gain +5 offspring	gain +5 offspring

Table 5.

As shown in Table 5, Smith constructed a payoff matrix in which various possible outcomes (*e.g.*, death, maiming, successful mating), and the costs and benefits associated with them (*e.g.*, cost of lost time), were weighted in terms of the expected number of genes propagated. Smith showed that a bourgeois invasion would be successful against a completely hawk population by observing that when a hawk confronts a hawk it loses 5, whereas a bourgeois loses only 2.5. (Because the population is assumed to be predominantly hawk, the success of the invasion can be predicted by comparing the average number of offspring a hawk will produce when it confronts another hawk with the average number of offspring a bourgeois will produce when confronting a hawk.) Patently, a bourgeois invasion against a completely dove population would be successful as well, gaining the bourgeois 6 offspring. On the other hand, a completely bourgeois population cannot be invaded by either hawks or doves, because the bourgeois gets 5 against bourgeois, which is more than either hawks or doves get when confronting bourgeois. Note in this application that the question is not what strategy a rational player will choose—animals are not assumed to make conscious choices, though their types may change through mutation—but what combinations of types are stable and hence likely to evolve.

Smith gave several examples of the bourgeois strategy. For example, male speckled wood butterflies seek sunlit spots on the forest floor where females are often found. There is a shortage of such spots, however, and in a confrontation between a stranger and an inhabitant, the stranger yields after a brief duel in which the combatants circle one another. The dueling skills of the adversaries have little effect on the outcome. When one butterfly is forcibly placed on another's territory so that each considers the other the aggressor, the two butterflies duel with righteous indignation for a much longer time.

N-PERSON GAMES

Nash equilibriums

Theoretically, *n*-person games in which the players are not allowed to communicate and make binding agreements are not fundamentally different from two-person noncooperative games. In the two examples that follow, each involving three players, one looks for Nash equilibriums—that is, stable outcomes from which no player would normally depart because to do so would be disadvantageous.

Sequential and simultaneous truels. As an example of an *n*-person noncooperative game, imagine three players, *A*, *B*, and *C*, situated at the corners of an equilateral triangle. They engage in a truel, or three-person duel, in which each player has a gun with one bullet. Assume that each player is a perfect shot and can kill one other player at any time. There is no fixed order of play, but any shooting that occurs is sequential: no player fires at the same time as any other. Consequently, if a bullet is fired, the results are known to all players before another bullet is fired.

Suppose that the players order their goals as follows: (1) survive alone, (2) survive with one opponent, (3) survive with both opponents, (4) not survive, with no opponents alive, (5) not survive, with one opponent alive, and (6) not survive, with both opponents alive. Thus, surviving alone is best, dying alone is worst.

If a player can either fire or not fire at another player, who, if anybody, will shoot whom? It is not difficult to see that outcome (3), in which nobody shoots, is the unique Nash equilibrium—any player that departs from not shooting does worse. Suppose, on the contrary, that *A* shoots *B*, hoping for *A*'s outcome (2), whereby he and *C* survive. Now, however, *C* can shoot a disarmed *A*, thereby leaving himself as the sole survivor, or outcome (1). As this is *A*'s penultimate outcome (5), in which *A* and one opponent (*B*) are killed while the other opponent (*C*) lives, *A* should not fire the first shot; the same reasoning applies to the other two players. Consequently, nobody will shoot, resulting in outcome (3), in which all three players survive.

Now consider whether any of the players can do better through collusion. Specifically, assume that *A* and *B* agree not to shoot each other; if either shoots another player, they agree it would be *C*. Nevertheless, if *A* shoots *C* (for instance), *B* could now repudiate the agreement with impunity and shoot *A*, thereby becoming the sole survivor.

Thus, thinking ahead about the consequences of shooting first or colluding with another player to do so, nobody will shoot or collude. Thereby all players will survive if the players must act in sequence, giving outcome (3). Because no player can do better by shooting, or saying they will do so to another, these strategies yield a Nash equilibrium.

Next, suppose that the players act simultaneously; hence, they must decide in ignorance of each others' intended actions. This situation is common in life: people often must act before they find out what others are doing. While there is no "best" strategy in all situations, the possibilities of survival will increase if the number of rounds is unlimited. Outcome: There may be zero, one, or three survivors, but never two. To summarize, shooting is never rational in a sequential truel, whereas it is always rational in a simultaneous truel that goes only one round. Thus, "nobody shoots" and "everybody shoots" are the Nash equilibriums in these two kinds of truels. In simultaneous truels that go more than one round, by comparison, there are multiple Nash equilibriums. If the number of rounds is known, then there is one Nash equilibrium in which a player shoots, and one in which he does not, at the start, but in the end there will be only one or no survivors. When the number of rounds is unlimited, however, a new Nash equilibrium is possible in which nobody shoots on any round. Thus, like PD with an uncertain number of rounds, an unlimited number of rounds in a truel can lead to greater cooperation.

The von Neumann–Morgenstern theory. Von Neumann and Morgenstern were the first to construct a cooperative theory of *n*-person games. They assumed that various groups of players might join together to form coalitions, each of which has an associated value defined as the minimum amount that the coalition can ensure by its own efforts. (In practice, such groups might be blocs in a legislative body or business partners in a conglomerate.) They described these *n*-person games in characteristic-function form—that is, by listing the individual players (one-person coalitions), all possible coalitions of two or more players, and the values that each of these coalitions could ensure if a counter-coalition comprising all other players acted to minimize the amount that the coalition can obtain. They also assumed that the characteristic function is superadditive: the value of a coalition of two formerly separate coalitions is at least as great as the sum of the separate values of the two coalitions.

The sum of payments to the players in each coalition must equal the value of that coalition. Moreover, each player in a coalition must receive no less than what he could obtain playing alone; otherwise, he would not join the coalition. Each set of payments to the players describes one possible outcome of an *n*-person cooperative game and is called an imputation. Within a coalition *S*, an imputation *X* is said to dominate another imputation *Y* if each player in *S* gets more with *X* than with *Y* and if the players in *S* receive a total payment that does not exceed the coalition value of *S*. This means that players in the coalition prefer the payoff *X* to the payoff *Y* and have the power to enforce this preference.

Imputations

Von Neumann and Morgenstern defined the solution to an *n*-person game as a set of imputations satisfying two conditions: (1) no imputation in the solution dominates another imputation in the solution and (2) any imputation not in the solution is dominated by another one in the solution. A von Neumann–Morgenstern solution is not a single outcome but, rather, a set of outcomes, any one of which may occur. It is stable because, for the members of the coalition, any imputation outside the solution is dominated by—and is therefore less attractive than—an imputation within the solution. The imputations within the solution are viable because they are not dominated by any other imputations in the solution.

In any given cooperative game there are generally many—sometimes infinitely many—solutions. A simple three-person game that illustrates this fact is one in which any two players, as well as all three players, receive one unit, which they can divide between or among themselves in any way that they wish; individual players receive nothing. In such a case the value of each two-person coalition, and the three-person coalition as well, is 1.

One solution to this game consists of three imputations, in each of which one player receives 0 and the other two players receive ½ each. There is no self-domination within the solution, because if one imputation is substituted for another, one player gets more, one gets less, and one gets the same (for domination, each of the players forming a coalition must gain). In addition, any imputation outside the solution is dominated by one in the solution, because the two players with the lowest payoffs must each get less than ½; clearly, this imputation is dominated by an imputation in the solution in which these two players each get ½. According to this solution, at any given time one of its three imputations will occur, but von Neumann and Morgenstern do not predict which one.

A second solution to this game consists of all the imputations in which player *A* receives ¼ and players *B* and *C* share the remaining ¾. Although this solution gives a different set of outcomes from the first solution, it, too, satisfies von Neumann and Morgenstern's two conditions. For any imputation within the solution, player *A* always gets ¼ and therefore cannot gain. In addition, because players *B* and *C* share a fixed sum, if one of them gains in a proposed imputation, the other must lose. Thus, no imputation in the solution dominates another imputation in the solution.

For any imputation not in the solution, player *A* must get either more or less than ¼. When *A* gets more than ¼, players *B* and *C* share less than ¾ and, therefore, can do better with an imputation within the solution. When player *A* gets less than ¼, say ⅛, he always does better with an imputation in the solution. Players *B* and *C* now have more to share; but no matter how they split the new total of ⅞, there is an imputation in the solution that one of them will prefer. When they share equally, each gets 7/16; but player *B*, for example, can get more in the imputation (¼, ½, ¼), which is in the solution. When players *B* and *C* do not divide the ⅞ equally, the player who gets the smaller amount can always do better with an imputation in the solution. Thus, any imputation outside the solution is dominated by one inside the solution. Similarly, it can be shown that all of the imputations in which player *B* gets ¼ and players *A* and *C* share ¾, as well as the set of all imputations in which player *C* gets ¼ and players *A* and *B* share ¾, also constitute a solution to the game.

Although there may be many solutions to a game (each representing a different "standard of behaviour"), it was not apparent at first that there would always be at least one in every cooperative game. Von Neumann and Morgenstern found no game without a solution, and they deemed it important that no such game exists. However, in 1967 a fairly complicated 10-person game was discovered by the American mathematician William F. Lucas that did not have a solution. This and later counterexamples indicated that the von Neumann–Morgenstern solution is not universally applicable, but it remains compelling, especially since no definitive theory of *n*-person cooperative games exists.

Game without a solution

The Banzhaf value in voting games. Power defined as control over outcomes is not synonymous with control over resources. For example, in the paradox of the chair's position, the chair's extra tie-breaking vote may actually be a detriment because the strategic situation facing voters may intervene and cause them to reassess their strategies in light of the additional resources that the chair possesses. In

Paradox of the chair

doing so, they may be led to "gang up" against the chair. In effect, the chair's resources become a burden to bear, not power to relish.

When players' preferences are not known beforehand, though, it is useful to define power in terms of their ability to alter the outcome by changing their votes, as governed by a constitution, bylaws, or other rules of the game. Various measures of voting power have been proposed for simple games, in which every coalition has a value of 1 (if it has enough votes to win) or 0 (if it does not). The sum of the powers of all the players is 1. When a player has 0 power, his vote has no influence on the outcome; when a player has a power of 1, the outcome depends only on his vote. The key to calculating voting power is determining the frequency with which a player casts a critical vote.

Critical vote

American attorney John F. Banzhaf III proposed that all combinations in which any player is the critical voter—that is, in which a measure passes only with this voter's support—be considered equally likely. The Banzhaf value for each player is then the number of combinations in which this voter is critical divided by the total number of combinations in which each voter (including this one) is critical.

This view is not compatible with defining the voting power of a player to be proportional to the number of votes he casts, because votes per se may have little or no bearing on the choice of outcomes. For example, in a three-member voting body in which *A* has 4 votes, *B* 2 votes, and *C* 1 vote, members *B* and *C* will be powerless if a simple majority wins. The fact that members *B* and *C* together control $3/7$ of the votes is irrelevant in the selection of outcomes, so these members are called dummies. Member *A*, by contrast, is a dictator by virtue of having enough votes alone to determine the outcome. A voting body can have only one dictator, whose existence renders all other members dummies, but there may be dummies and no dictator.

Minimal winning coalition

A minimal winning coalition (MWC) is one in which the subtraction of at least one of its members renders it losing. To illustrate the calculation of Banzhaf values, consider a voting body with two 2-vote members (distinguished as 2a and 2b) and one 3-vote member, in which a simple majority wins. There are three distinct MWCs—(3, 2a), (3, 2b), and (2a, 2b)—or combinations in which some voter is critical; the grand coalition, comprising all three members, (3, 2a, 2b), is not an MWC because no single member's defection would cause it to lose.

As each member's defection is critical in two MWCs, each member's proportion of voting power is two-sixths, or one-third. Thus, the Banzhaf index, which gives the Banzhaf values for each member in vector form, is (⅓, ⅓, ⅓). Clearly, the voting power of the 3-vote member is the same as that of each of the two 2-vote members, although the 3-vote member has 50 percent greater weight (more votes) than each of the 2-vote members.

The discrepancy between voting weight and voting power is more dramatic in the voting body (50, 49, 1) where, again, a simple majority wins. The 50-vote member is critical in all three MWCs—(50, 1), (50, 49), and (50, 49, 1), giving him a veto because his presence is necessary for a coalition to be winning—whereas the 49-vote member is critical in only (50, 49) and the 1-vote member in only (50, 1). Thus, the Banzhaf index for (50, 49, 1) is (⅗, ⅕, ⅕), making the 49-vote member indistinguishable from the 1-vote member; the 50-vote member, with just one more vote than the 49-vote member, has three times as much voting power.

In 1958 six western European countries formed the European Economic Community (EEC). The three large countries (West Germany, France, and Italy) each had 4 votes on its Council of Ministers, the two medium-size countries (Belgium and The Netherlands) 2 votes each, and the one small country (Luxembourg) 1 vote. The decision rule of the Council was a qualified majority of 12 out of 17 votes, giving the large countries Banzhaf values of $5/21$ each, the medium-size countries ⅐ each, and—amazingly—Luxembourg no voting power at all. From 1958 to 1973—when the EEC admitted three additional members—Luxembourg was a dummy. Luxembourg might as well not have gone to Council meetings except to participate in the debate, because its 1 vote could never change the outcome. To see this without calculating the Banzhaf values of all the members, note that the votes of the five other countries are all even numbers. Therefore, an MWC with exactly 12 votes could never include Luxembourg's (odd) 1 vote; while a 13-vote MWC that included Luxembourg could form, Luxembourg's defection would never render such an MWC losing. It is worth noting that as the Council kept expanding with the addition of new countries and the formation of the European Union, Luxembourg never reverted to being a dummy, even though its votes became an ever smaller proportion of the total.

The Banzhaf and other power indices, rooted in cooperative game theory, have been applied to many voting bodies, not necessarily weighted, sometimes with surprising results. For example, the Banzhaf index has been used to calculate the power of the 5 permanent and 10 nonpermanent members of the United Nations Security Council. (The permanent members, all with a veto, have 83 percent of the power.) It has also been used to compare the power of representatives, senators, and the president in the U.S. federal system.

Game theory is now well established and widely used in a variety of disciplines. The foundations of economics, for example, are increasingly grounded in game theory; among game theory's many applications in economics is the design of Federal Communications Commission auctions of airwaves, which have netted the U.S. government billions of dollars. Game theory is being used increasingly in political science to study strategy in areas as diverse as campaigns and elections, defense policy, and international relations. In biology, business, management science, computer science, and law, game theory has been used to model a variety of strategic situations. Game theory has even penetrated areas of philosophy (*e.g.*, to study the equilibrium properties of ethical rules), religion (*e.g.*, to interpret Bible stories), and pure mathematics (*e.g.*, to analyze how to divide a cake fairly among *n* people). All in all, game theory holds out great promise not only for advancing the understanding of strategic interaction in very different settings but also for offering prescriptions for the design of better auction, bargaining, voting, and information systems that involve strategic choice. (M.D.D./S.J.B.)

BIBLIOGRAPHY. The seminal work in game theory is JOHN VON NEUMANN and OSKAR MORGENSTERN, *Theory of Games and Economic Behavior*, 3rd ed. (1953, reprinted 1980). AVINASH K. DIXIT and BARRY J. NALEBUFF, *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life* (1991), uses case studies, without formal mathematical analysis, to introduce the principles of game theory. Two introductions that require only high school algebra are AVINASH K. DIXIT and SUSAN SKEATH, *Games of Strategy* (1999); and PHILIP D. STRAFFIN, *Game Theory* (1993).

Applications of game theory are presented in NESMITH C. ANKENY, *Poker Strategy: Winning with Game Theory* (1981, reprinted 1982); ROBERT AXELROD, *The Evolution of Cooperation* (1984); DOUGLAS G. BAIRD, ROBERT H. GERTNER, and RANDAL C. PICKER, *Game Theory and the Law* (1994); STEVEN J. BRAMS, *Biblical Games: Game Theory and the Hebrew Bible*, 2nd ed. (2002); STEVEN J. BRAMS, *Theory of Moves* (1994); DAN S. FELSENTHAL and MOSHÉ MACHOVER, *The Measurement of Voting Power: Theory and Practice, Problems and Paradoxes* (1998); BARRY O'NEILL, *Honor, Symbols, and War* (1999); and KARL SIGMUND, *Games of Life: Explorations in Ecology, Evolution, and Behavior* (1993).

Histories of game theory can be found in WILLIAM POUNDSTONE, *Prisoner's Dilemma: John von Neumann, Game Theory, and the Puzzle of the Bomb* (1992); and E. ROY WEINTRAUB (ed.), *Toward a History of Game Theory* (1992). (S.J.B.)

Gandhi

Mohandas Karamchand Gandhi, the preeminent leader of Indian nationalism and the prophet of nonviolence in the 20th century, was born, the youngest child of his father's fourth wife, on Oct. 2, 1869, at Porbandar, the capital of a small principality in Gujarāt in western India under British suzerainty. His father, Karamchand Gandhi, who was the dewan (chief minister) of Porbandar, did not have much in the way of a formal education but was an able administrator who knew how to steer his way between the capricious princes, their long-suffering subjects, and the headstrong British political officers in power.

Gandhi's mother, Putlibai, was completely absorbed in religion, did not care much for finery and jewelry, divided her time between her home and the temple, fasted frequently, and wore herself out in days and nights of nursing whenever there was sickness in the family. Mohandas grew up in a home steeped in Vaishnavism (Vaiṣṇavism)—worship of the Hindu god Vishnu (Viṣṇu)—with a strong tinge of Jainism, a morally rigorous Indian religion, whose chief tenets are nonviolence and the belief that everything in the universe is eternal. Thus he took for granted *ahiṁsā* (noninjury to all living beings), vegetarianism, fasting for self-purification, and mutual tolerance between adherents of various creeds and sects.

Youth. The educational facilities at Porbandar were rudimentary; in the primary school that Mohandas attended, the children wrote the alphabet in the dust with their fingers. Luckily for him, his father became dewan of Rājkot, another princely state. Though he occasionally won prizes and scholarships at the local schools, his record was on the whole mediocre. One of the terminal reports rated him as "good at English, fair in Arithmetic and weak in Geography; conduct very good, bad handwriting." A diffident child, he was married at the age of 13 and thus lost a year at school. He shone neither in the classroom nor on the playing field. He loved to go out on long solitary walks when he was not nursing his by now ailing father or helping his mother with her household chores.

Education

He had learned, in his words, "to carry out the orders of the elders, not to scan them." With such extreme passivity, it is not surprising that he should have gone through a phase of adolescent rebellion, marked by secret atheism, petty thefts, furtive smoking, and—most shocking of all for a boy born in a Vaishnava family—meat eating. His adolescence was probably no stormier than that of most children of his age and class. What was extraordinary was the way his youthful transgressions ended.

Margaret Bourke-White, LIFE MAGAZINE © TIME INC

Gandhi, 1946.

"Never again" was his promise to himself after each escapade. And he kept his promise. Beneath an unprepossessing exterior, he concealed a burning passion for self-improvement that led him to take even the heroes of Hindu mythology, such as Prahlāda and Hariścandra—legendary embodiments of truthfulness and sacrifice—as living models.

In 1887 Mohandas scraped through the matriculation examination of the University of Bombay and joined Samaldas College in Bhāvnagar (Bhaunagar). As he had suddenly to switch from his native language—Gujarātī—to English, he found it rather difficult to follow the lectures.

Meanwhile, his family was debating his future. Left to himself, he would have liked to be a doctor. But, besides the Vaishnava prejudice against vivisection, it was clear that, if he was to keep up the family tradition of holding high office in one of the states in Gujarāt, he would have to qualify as a barrister. This meant a visit to England, and Mohandas, who was not too happy at Samaldas College, jumped at the proposal. His youthful imagination conceived England as "a land of philosophers and poets, the very centre of civilization." But there were several hurdles to be crossed before the visit to England could be realized. His father had left little property; moreover, his mother was reluctant to expose her youngest child to unknown temptations and dangers in a distant land. But Mohandas was determined to visit England. One of his brothers raised the necessary money, and his mother's doubts were allayed when he took a vow that, while away from home, he would not touch wine, women, or meat. Mohandas disregarded the last obstacle—the decree of the leaders of the Modh Bania subcaste (Vaisya caste), to which the Gandhis belonged, who forbade his trip to England as a violation of the Hindu religion—and sailed in September 1888. Ten days after his arrival, he joined the Inner Temple, one of the four London law colleges.

Law studies

England. Gandhi took his studies seriously and tried to brush up on his English and Latin by taking the London University matriculation examination. But, during the three years he spent in England, his main preoccupation was with personal and moral issues rather than with academic ambitions. The transition from the half-rural atmosphere of Rājkot to the cosmopolitan life of London was not easy for him. As he struggled painfully to adapt himself to Western food, dress, and etiquette, he felt awkward. His vegetarianism became a continual source of embarrassment to him; his friends warned him that it would wreck his studies as well as his health. Fortunately for him he came across a vegetarian restaurant as well as a book providing a reasoned defense of vegetarianism, which henceforth became a matter of conviction for him, not merely a legacy of his Vaishnava background. The missionary zeal he developed for vegetarianism helped to draw the pitifully shy youth out of his shell and gave him a new poise. He became a member of the executive committee of the London Vegetarian Society, attending its conferences and contributing articles to its journal.

In the vegetarian restaurants and boarding houses of England, Gandhi met not only food faddists but some earnest men and women to whom he owed his introduction to the Bible and the *Bhagavadgītā*, the most popular expression of Hinduism in the form of a philosophical poem, which he read for the first time in its English translation by Sir Edwin Arnold. The English vegetarians were a motley crowd. They included socialists and humanitarians like Edward Carpenter, "the British Thoreau"; Fabians like George Bernard Shaw; and Theosophists like Annie Besant. Most of them were idealists; quite a few were rebels who rejected the prevailing values of the late Victorian

Establishment, denounced the evils of the capitalist and industrial society, preached the cult of the simple life, and stressed the superiority of moral over material values and of cooperation over conflict. These ideas were to contribute substantially to the shaping of Gandhi's personality and, eventually, to his politics.

Return to India

Painful surprises were in store for Gandhi when he returned to India in July 1891. His mother had died in his absence, and he discovered to his dismay that the barrister's degree was not a guarantee of a lucrative career. The legal profession was already beginning to be overcrowded, and Gandhi was much too diffident to elbow his way into it. In the very first brief he argued in a Bombay court, he cut a sorry figure. Turned down even for the part-time job of a teacher in a Bombay high school, he returned to Rājkot to make a modest living by drafting petitions for litigants. Even this employment was closed to him when he incurred the displeasure of a local British officer. It was, therefore, with some relief that he accepted the none-too-attractive offer of a year's contract from an Indian firm in Natal, South Africa.

Reaction to segregation

South Africa. Africa was to present to Gandhi challenges and opportunities that he could hardly have conceived. In a Durban court, he was asked by the European magistrate to take off his turban; he refused and left the courtroom. A few days later, while travelling to Pretoria, he was unceremoniously thrown out of a first-class railway compartment and left shivering and brooding at Pietermaritzburg Station; in the further course of the journey he was beaten up by the white driver of a stagecoach because he would not travel on the footboard to make room for a European passenger; and finally he was barred from hotels reserved "for Europeans only." These humiliations were the daily lot of Indian traders and labourers in Natal who had learned to pocket them with the same resignation with which they pocketed their meagre earnings. What was new was not Gandhi's experience but his reaction. He had so far not been conspicuous for self-assertion or aggressiveness. But something happened to him as he smarted under the insults heaped upon him. In retrospect the journey from Durban to Pretoria struck him as one of the most creative experiences of his life; it was his moment of truth. Henceforth he would not accept injustice as part of the natural or unnatural order in South Africa; he would defend his dignity as an Indian and as a man.

While in Pretoria, Gandhi studied the conditions in which his countrymen lived and tried to educate them on their rights and duties, but he had no intention of staying on in South Africa. Indeed, in June 1894, as his year's contract drew to a close, he was back in Durban, ready to sail for India. At a farewell party given in his honour he happened to glance through the *Natal Mercury* and learned that the Natal Legislative Assembly was considering a bill to deprive Indians of the right to vote. "This is the first nail in our coffin," Gandhi told his hosts. They professed their inability to oppose the bill, and indeed their ignorance of the politics of the colony, and begged him to take up the fight on their behalf.

Until the age of 18, Gandhi had hardly ever read a newspaper. Neither as a student in England nor as a budding barrister in India had he evinced much interest in politics. Indeed, he was overcome by a terrifying stage fright whenever he stood up to read a speech at a social gathering or to defend a client in court. Nevertheless, in July 1894, when he was barely 25, he blossomed almost overnight into a proficient political campaigner. He drafted petitions to the Natal legislature and the British government and had them signed by hundreds of his compatriots. He could not prevent the passage of the bill but succeeded in drawing the attention of the public and the press in Natal, India, and England to the Natal Indians' grievances. He was persuaded to settle down in Durban to practice law and to organize the Indian community. In 1894, he founded the Natal Indian Congress of which he himself became the indefatigable secretary. Through this common political organization, he infused a spirit of solidarity in the heterogeneous Indian community. He flooded the government, the legislature, and the press with closely reasoned statements of Indian grievances. Finally, he exposed to the view of the outside world the skeleton in the imperial cupboard, the discrimination practiced against the Indian subjects of Queen Victoria in one of her own colonies in Africa. It was a measure of his success as a publicist that such important newspapers as *The Times* of London and the *Statesman* and *Englishman* of Calcutta editorially commented on the Natal Indians' grievances.

In 1896 Gandhi went to India to fetch his wife Kasturbai and their children and to canvass support for the Indians overseas. He met prominent leaders and persuaded them to address public meetings in the country's principal cities. Unfortunately for him, garbled versions of his activities and utterances reached Natal and inflamed its European population. On landing at Durban in January 1897, he was assaulted and nearly lynched by a white mob. Joseph Chamberlain, the colonial secretary in the British Cabinet, cabled the government of Natal to bring the guilty men to book, but Gandhi refused to prosecute his assailants. It was, he said, a principle with him not to seek redress of a personal wrong in a court of law.

Role in the Boer War

Gandhi was not the man to nurse a grudge. On the outbreak of the Boer War in 1899, he argued that the Indians, who claimed the full rights of citizenship in the British crown colony of Natal, were in duty bound to defend it. He raised an ambulance corps of 1,100 volunteers, out of whom 300 were free Indians and the rest indentured labourers. It was a motley crowd: barristers and accountants, artisans and labourers. It was Gandhi's task to instill in them a spirit of service to those whom they regarded as their oppressors. The editor of the *Pretoria News* has left a fascinating pen portrait of Gandhi in the battle zone:

> After a night's work which had shattered men with much bigger frames, I came across Gandhi in the early morning sitting by the roadside eating a regulation army biscuit. Every man in (General) Buller's force was dull and depressed, and damnation was heartily invoked on everything. But Gandhi was stoical in his bearing, cheerful and confident in his conversation and had a kindly eye.

The British victory in the Boer War brought little relief to the Indians in South Africa. The new regime in South Africa was to blossom into a partnership, but only between Boers and Britons. Gandhi saw that, with the exception of a few Christian missionaries and youthful idealists, he had been unable to make a perceptible impression upon the South African Europeans. In 1906 the Transvaal government published a particularly humiliating ordinance for the registration of its Indian population. The Indians held a mass protest meeting at Johannesburg in September 1906 and, under Gandhi's leadership, took a pledge to defy the ordinance if it became law in the teeth of their opposition, and to suffer all the penalties resulting from their defiance. Thus was born *satyāgraha* ("devotion to truth"), a new technique for redressing wrongs through inviting, rather than inflicting, suffering, for resisting the adversary without rancour and fighting him without violence.

The struggle in South Africa lasted for more than seven years. It had its ups and downs, but under Gandhi's leadership, the small Indian minority kept up its resistance against heavy odds. Hundreds of Indians chose to sacrifice their livelihood and liberty rather than submit to laws repugnant to their conscience and self-respect. In the final phase of the movement in 1913, hundreds of Indians, including women, went to jail, and thousands of Indian workers who had struck work in the mines bravely faced imprisonment, flogging, and even shooting. It was a terrible ordeal for the Indians, but it was also the worst possible advertisement for the South African government, which, under pressure from the governments of Britain and India, accepted a compromise negotiated by Gandhi on the one hand and the South African statesman General Jan Christian Smuts on the other.

"The saint has left our shores," Smuts wrote to a friend on Gandhi's departure from South Africa for India, in July 1914, "I hope for ever." Twenty-five years later, he wrote that it had been his "fate to be the antagonist of a man for whom even then I had the highest respect." Once, during his not infrequent stays in jail, Gandhi had prepared a pair of sandals for Smuts, who recalled that there was no hatred and personal ill-feeling between them, and when the fight was over "there was the atmosphere in which a decent peace could be concluded."

As later events were to show, Gandhi's work did not provide an enduring solution for the Indian problem in South Africa. What he did to South Africa was indeed less important than what South Africa did to him. It had not treated him kindly, but, by drawing him into the vortex of its racial problem, it had provided him with the ideal setting in which his peculiar talents could unfold themselves.

The religious quest. Gandhi's religious quest dated back to his childhood, the influence of his mother and of his home at Porbandar and Rājkot, but it received a great impetus after his arrival in South Africa. His Quaker friends in Pretoria failed to convert him to Christianity, but they quickened his appetite for religious studies. He was fascinated by Tolstoy's writings on Christianity, read the Qu'rān in translation, and delved into Hindu scriptures and philosophy. The study of comparative religion, talks with scholars, and his own reading of theological works brought him to the conclusion that all religions were true and yet every one of them was imperfect because they were "interpreted with poor intellects, sometimes with poor hearts, and more often misinterpreted."

Return to Hinduism

Rajchandra, a brilliant young philosopher who became Gandhi's spiritual mentor, convinced him of "the subtlety and profundity" of Hinduism, the religion of his birth. And it was the *Bhagavadgītā,* which Gandhi had first read in London, that became his "spiritual dictionary" and exercised probably the greatest single influence on his life. Two Sanskrit words in the *Gītā* particularly fascinated him. One was *aparigraha* (nonpossession), which implied that man had to jettison the material goods that cramped the life of the spirit and to shake off the bonds of money and property. The other was *samabhava* (equability), which enjoined him to remain unruffled by pain or pleasure, victory or defeat, and to work without hope of success or fear of failure.

These were not merely counsels of perfection. In the civil case that had brought him to South Africa in 1893, he had persuaded the antagonists to settle their differences out of court. The true function of a lawyer seemed to him "to unite parties riven asunder." He soon regarded his clients not as purchasers of his services but as friends; they consulted him not only on legal issues but on such matters as the best way of weaning a baby or balancing the family budget. When an associate protested that clients came even on Sundays, Gandhi replied: "A man in distress cannot have Sunday rest."

Gandhi's legal earnings reached a peak figure of £5,000 a year, but he had little interest in moneymaking, and his savings were often sunk in his public activities. In Durban and later in Johannesburg, he kept an open table; his house was a virtual hostel for younger colleagues and political coworkers. This was something of an ordeal for his wife, without whose extraordinary patience, endurance, and self-effacement Gandhi could hardly have devoted himself to public causes. As he broke through the conventional bonds of family and property, their life tended to shade into a community life.

Gandhi felt an irresistible attraction to a life of simplicity, manual labour, and austerity. In 1904, after reading John Ruskin's *Unto This Last,* a critique of capitalism, he set up a farm at Phoenix near Durban where he and his friends could literally live by the sweat of their brow. Six years later another colony grew up under Gandhi's fostering care near Johannesburg; it was named Tolstoy Farm after the Russian writer and moralist, whom Gandhi admired and corresponded with. Those two settlements were the precursors of the more famous ashrams (*āśramas*) in India, at Sābarmati near Ahmedabad (Ahmadābād) and at Sevagram near Wardha.

South Africa had not only prompted Gandhi to evolve a novel technique for political action but also transformed him into a leader of men by freeing him from bonds that make cowards of most men. "Persons in power," Gilbert Murray prophetically wrote about Gandhi in the *Hibbert Journal* in 1918, "should be very careful how they deal with a man who cares nothing for sensual pleasure, nothing for riches, nothing for comfort or praise, or promotion, but is simply determined to do what he believes to be right. He is a dangerous and uncomfortable enemy, because his body which you can always conquer gives you so little purchase upon his soul."

Emergence as leader of nationalist India. From 1915 to 1918, Gandhi seemed to hover uncertainly on the periphery of Indian politics, declining to join any political agitation, supporting the British war effort in World War I, and even recruiting soldiers for the British Indian Army. At the same time, he did not flinch from criticizing the British officials for any acts of high-handedness or from taking up the grievances of the long-suffering peasantry in Bihār and Gujarāt. Not until February 1919, provoked by the British insistence on pushing through the Rowlatt Bills, which empowered the authorities to imprison without trial those suspected of sedition, in the teeth of Indian opposition, did Gandhi reveal a sense of estrangement from the British Raj. He announced a *satyāgraha* struggle. The result was a virtual political earthquake that shook the subcontinent in the spring of 1919. The violent outbreaks that followed—leading, among other incidents, to the killing by British-led soldiers of nearly 400 Indians attending a meeting at Amritsar in the Punjab and the enactment of martial law—prompted him to stay his hand. But within a year he was again in a militant mood, having in the meantime been irrevocably alienated by British insensitiveness to Indian feeling on the Punjab tragedy and Muslim resentment on the peace terms offered to Turkey following World War I.

Massacre at Amritsar

By the autumn of 1920, Gandhi was the dominant figure on the political stage, commanding an influence never attained by any political leader in India or perhaps in any other country. He refashioned the 35-year-old Indian National Congress into an effective political instrument of Indian nationalism: from a three-day Christmas-week picnic of the upper middle class in one of the principal cities of India, it became a mass organization with its roots in small towns and villages. Gandhi's message was simple; it was not British guns but imperfections of Indians themselves that kept their country in bondage. His program of nonviolent noncooperation with the British government included boycott not only of British manufactures but of institutions operated or aided by the British in India: legislatures, courts, offices, schools. This program electrified the country, broke the spell of fear of foreign rule, and led to arrests of thousands of *satyāgrahi*s, who defied laws and cheerfully lined up for prison. In February 1922 the movement seemed to be on the crest of a rising wave, but, alarmed by a violent outbreak in Chauri Chaura, a remote village in eastern India, Gandhi decided to call off mass civil disobedience. This was a blow to many of his followers, who feared that his self-imposed restraints and scruples would reduce the nationalist struggle to pious futility. Gandhi himself was arrested on March 10, 1922, tried for sedition, and sentenced to six years' imprisonment. He was released in February 1924, after an operation for appendicitis. The political landscape had changed in his absence. The Congress Party had split into two factions, one under Chitta Ranjan Das and Motilal Nehru (the father of Jawaharlal Nehru, India's first prime minister) favouring the entry of the party into legislatures and the other under C. Rajagopalachari and Vallabhbhai Jhaverbhai Patel opposing it. Worst of all, the unity between Hindus and Muslims of the heyday of the noncooperation movement of 1920–22 had dissolved. Gandhi tried to draw the warring communities out of their suspicion and fanaticism by reasoning and persuasion. And finally, after a serious communal outbreak, he undertook a three-week fast in the autumn of 1924 to arouse the people into following the path of nonviolence.

During the mid-1920s Gandhi took little interest in active politics and was considered a spent force. But in 1927 the British government appointed a constitutional reform commission under Sir John Simon, a prominent English lawyer and politician, that did not contain a single Indian. When the Congress and other parties boycotted the commission, the political tempo rose. After the Calcutta Congress in December 1928, where Gandhi moved the crucial resolution demanding dominion status from

Return to party leadership

the British government within a year under threat of a nation-wide nonviolent campaign for complete independence, Gandhi was back at the helm of the Congress Party. In March 1930, he launched the *satyāgraha* against the tax on salt, which affected the poorest section of the community. One of the most spectacular and successful campaigns in Gandhi's nonviolent war against the British Raj, it resulted in the imprisonment of more than 60,000 persons. A year later, after talks with Lord Irwin, Gandhi accepted a truce, called off civil disobedience, and agreed to attend the Round Table Conference in London as the sole representative of the Indian National Congress. The conference, which concentrated on the problem of the Indian minorities rather than on the transfer of power from the British, was a great disappointment to the Indian nationalists. Moreover, when Gandhi returned to India in December 1931 he found his party facing an all-out offensive from Lord Irwin's successor, Lord Willingdon, who unleashed the sternest repression in the history of the nationalist movement. Gandhi was once more imprisoned, and the government tried to insulate him from the outside world and to destroy his influence. This was not an easy task. Gandhi soon regained the initiative; in September 1932, while still a prisoner, he embarked on a fast to protest against the British government's decision to segregate the untouchables (the depressed classes) by allotting them separate electorates in the new constitution. The fast produced an emotional upheaval in the country; an alternative electoral arrangement was jointly and speedily devised by the leaders of the Hindu community and the untouchables and endorsed by the British government. The fast became the starting point of a vigorous campaign for the removal of the disabilities of the untouchables whom Gandhi renamed Harijans, "the children of God."

In 1934 Gandhi resigned not only as the leader but also as a member of the Congress Party. He had come to believe that its leading members had adopted nonviolence as a political expedient and not as the fundamental creed it was for him. In place of political activity he now concentrated on his "constructive programme" of building the nation "from the bottom up"—educating rural India, which accounted for 85 percent of the population; continuing his fight against untouchability; promoting handspinning, weaving, and other cottage industries to supplement the earnings of the underemployed peasantry; and evolving a system of education best suited to the needs of the people. Gandhi himself went to live at Sevagram, a village in central India, which became the centre of his program of social and economic uplift.

The last phase. With the outbreak of World War II, the nationalist struggle in India entered its last crucial phase. Gandhi hated fascism and all it stood for, but he also hated war. The Indian National Congress, on the other hand, was not committed to pacifism and was prepared to support the British war effort if Indian self-government was assured. Once more Gandhi became politically active. The failure of the mission of Sir Stafford Cripps, a British cabinet minister, who came to India in March 1942 with an offer that Gandhi found unacceptable, the British equivocation on the transfer of power to Indian hands, and the encouragement given by high British officials to conservative and communal forces promoting discord between Muslims and Hindus impelled him to demand in the summer of 1942 an immediate British withdrawal from India. The war against the Axis, particularly Japan, was in a critical phase; the British reacted sharply by imprisoning the entire Congress leadership and set out to crush the party once and for all. There were violent outbreaks that were sternly suppressed; the gulf between Britain and India became wider than ever.

A new chapter in Indo-British relations opened with the victory of the Labour Party in 1945. During the next two years, there were prolonged triangular negotiations between leaders of the Congress and the Muslim League under M.A. Jinnah and the British government culminating in the Mountbatten Plan of June 3, 1947, and the formation of the two new dominions of India and Pakistan in mid-August 1947.

Partition of British India

It was one of the greatest disappointments of Gandhi's life that Indian freedom was realized without Indian unity. Muslim separatism had received a great boost while Gandhi and his colleagues were in jail, and in 1946–47, as the final constitutional arrangements were being negotiated, the outbreak of communal riots between Hindus and Muslims unhappily created a climate in which Gandhi's appeals to reason and justice, tolerance and trust had little chance. When partition of the subcontinent was accepted—against his advice—he threw himself heart and soul into the task of healing the scars of the communal conflict, toured the riot-torn areas in Bengal and Bihār, admonished the bigots, consoled the victims, and tried to rehabilitate the refugees. In the atmosphere of that period, surcharged with suspicion and hatred, this was a difficult and heartbreaking task. Gandhi was blamed by partisans of both the communities. When persuasion failed, he went on a fast. He won at least two spectacular triumphs; in September 1947 his fasting stopped the rioting in Calcutta, and in January 1948, he shamed the city of Delhi into a communal truce. A few days later, on January 30, while he was on his way to his evening prayer meeting in Delhi, he was shot down by Nathuram Godse, a young Hindu fanatic.

Place in history. The British attitude to Gandhi was one of mingled admiration, amusement, bewilderment, suspicion, and resentment. Except for a tiny minority of Christian missionaries and radical socialists, the British tended to see in him at best a utopian visionary, at worst a cunning hypocrite whose professions of friendship for the British race were a mask for subversion of the British Raj. Gandhi was conscious of the existence of this wall of prejudice, and it was part of the strategy of *satyāgraha* to penetrate it.

His three major campaigns in 1920–22, 1930–34, and 1940–42 were well designed to engender that process of self-doubt and questioning that was to undermine the moral defences of his adversaries and to contribute, together with the objective realities of the postwar world, to producing the grant of dominion status in 1947. The British abdication in India was the first step in the liquidation of the British Empire on the continents of Asia and Africa. Gandhi's image as an archrebel died hard, but, as it had done to the memory of George Washington, Britain, in 1969, the centenary year of Gandhi's birth, erected a statue to his memory.

Gandhi had critics in his own country, and indeed in his own party. The liberal leaders protested that he was going too fast; the young radicals complained that he was not going fast enough; left-wing politicians alleged that he was not serious about evicting the British or liquidating such vested Indian interests as princes and landlords; the leaders of the untouchables doubted his good faith as a social reformer; and Muslim leaders accused him of partiality to his own community.

Recent research has established Gandhi's role as a great mediator and reconciler. His talents in this direction were applied to conflicts between the older moderate politicians and the young radicals, the political terrorists and the parliamentarians, the urban intelligentsia and the rural masses, the traditionalists and the modernists, the caste Hindus and the untouchables, the Hindus and the Muslims, and the Indians and the British.

Primacy of religion

It was inevitable that Gandhi's role as a political leader should loom larger in public imagination, but the mainspring of his life lay in religion, not in politics. And religion for him did not mean formalism, dogma, ritual, or sectarianism. "What I have been striving and pining to achieve these thirty years," he wrote in his autobiography, "is to see God face to face." His deepest strivings were spiritual, but unlike many of his countrymen with such aspirations, he did not retire to a cave in the Himalayas to meditate on the Absolute; he carried his cave, as he once said, within him. For him truth was not something to be discovered in the privacy of one's personal life; it had to be upheld in the challenging contexts of social and political life.

In the eyes of millions of his countrymen, he was the Mahatma (the great soul). The unthinking adoration of the huge crowds that gathered to see him all along his route made his tours a severe ordeal; he could hardly work during the day or rest at night. "The woes of the Mahatmas," he wrote, "are known only to the Mahatmas."

Gandhi won the affection and loyalty of gifted men and women, old and young, with vastly dissimilar talents and temperaments; of Europeans of every religious persuasion; and of Indians of almost every political line. Few of his political colleagues went all the way with him and accepted nonviolence as a creed; fewer still shared his food fads, his interest in mudpacks and nature cure, or his prescription of *brahmacarya,* complete renunciation of the pleasures of the flesh.

Gandhi's ideas on sex may sound quaint and unscientific. His marriage at the age of 13 seems to have complicated his attitude to sex and charged it with feelings of guilt, but it is important to remember that total sublimation, according to the best tradition of Hindu thought, is indispensable for those who seek self-realization, and *brahmacarya* was for Gandhi part of a larger discipline in food, sleep, thought, prayer, and daily activity designed to equip himself for service of the causes to which he was totally committed. What he failed to see was that his own unique experience was no guide for the common man.

It is probably too early to judge Gandhi's place in history. He was the catalyst if not the initiator of three of the major revolutions of the 20th century: the revolutions against colonialism, racism, and violence. He wrote copiously; the collected edition of his writings runs to more than 80 volumes.

Much of what he wrote was in response to the needs of his co-workers and disciples and the exigencies of the political situation, but on fundamentals, he maintained a remarkable consistency, as is evident from the *Hind Swaraj* ("Indian Home Rule") published in South Africa in 1909. The strictures on Western materialism and colonialism, the reservations about industrialism and urbanization, the distrust of the modern state, and the total rejection of violence that was expressed in this book seemed romantic, if not reactionary, to the pre-World War I generation in India and the West, which had not known the shocks of two global wars, experienced the phenomenon of Hitler, and the trauma of the atom bomb. Prime Minister Jawaharlal Nehru's objective of promoting a just and egalitarian order at home, and nonalignment with military blocs abroad doubtless owed much to Gandhi, but neither he nor his colleagues in the Indian nationalist movement wholly accepted the Gandhian models in politics and economics.

In recent years Gandhi's name has been invoked by the organizers of numerous demonstrations and movements, but with a few outstanding exceptions—such as those of his disciple the land reformer Vinoba Bhave in India and the black civil rights leader Martin Luther King, Jr., in the United States—these movements have been a travesty of the ideas of Gandhi.

Yet Gandhi will probably never lack champions. Erik H. Erikson, a distinguished American psychoanalyst, in his study of Gandhi senses "an affinity between Gandhi's truth and the insights of modern psychology." One of the greatest admirers of Gandhi was Albert Einstein, who saw in Gandhi's nonviolence a possible antidote to the massive violence unleashed by the fission of the atom. And Gunnar Myrdal, the Swedish economist, after his survey of the socioeconomic problems of the underdeveloped world, pronounced Gandhi "in practically all fields an enlightened liberal." In a time of deepening crisis in the underdeveloped world, of social malaise in the affluent societies, of the shadow of unbridled technology and the precarious peace of nuclear terror, it seems likely that Gandhi's ideas and techniques will become increasingly relevant. (B.R.N.)

BIBLIOGRAPHY. Gandhi's autobiography, *The Story of My Experiments with Truth,* 2 vol. (1927–29, reissued in 1 vol., 1983), tells the story of his life up to 1921; his *Satyagraha in South Africa,* 2nd ed. (1950, reprinted 1972), illuminates the formative two decades he spent in South Africa. The *Collected Works of Mahatma Gandhi,* 90 vol. (1958–84), includes all his writings, speeches, and letters.

A biography by PYARELAL, *Mahatma Gandhi,* 2nd ed., 2 vol. (1965–66), provides a richly documented chronicle of Gandhi's early and last years written by his former secretary. SUDHIR GHOSH, *Gandhi's Emissary* (1967), is an autobiographical memoir of Gandhi's informal agent to the British government in 1945–48. DINANATH G. TENDULKAR, *Mahatma,* rev. ed., 8 vol. (1960–63, reprinted 1969), tells the story of Gandhi's life mostly in Gandhi's own words extracted from his published writings. LOUIS FISCHER, *The Life of Mahatma Gandhi* (1950, reissued 1983), is based largely on printed sources but includes the author's vivid personal impressions of Gandhi and India in the 1940s; BAL R. NANDA, *Mahatma Gandhi: A Biography* (1958, reissued 1968), is a story of Gandhi's life as well as a critique of his thought and makes use of unpublished government records and correspondence of Gandhi. PENDEREL MOON, *Gandhi and Modern India* (1969), reflects a British administrator's views on Gandhi the politician. HENRY S.L. POLAK, HENRY M. BRAILSFORD, and FREDERICK W. PETHICK-LAWRENCE, *Mahatma Gandhi* (1949, reissued 1962), is a good introduction for Western readers. HORACE ALEXANDER, *Gandhi Through Western Eyes* (1969); and GEOFFREY ASHE, *Gandhi: A Study in Revolution* (1968), are sympathetic and analytical studies. ROBERT PAYNE, *The Life and Death of Mahatma Gandhi* (1969), is a well-researched biography, with emphasis on the personal rather than political aspect.

CHANDRAN D.S. DEVANESEN, *The Making of the Mahatma* (1969), covers Gandhi's childhood and youth in detail. ERIK H. ERIKSON, *Gandhi's Truth: On the Origins of Militant Nonviolence* (1969), illuminates Gandhi's life and technique by bringing to bear on them the insights of psychoanalysis. Another psychological biography is E. VICTOR WOLFENSTEIN, *The Revolutionary Personality: Lenin, Trotsky, Gandhi* (1967, reprinted 1971). See also JOSEPH J. DOKE, *M.K. Gandhi: An Indian Patriot in South Africa* (1909, reprinted 1967); CALVIN KYTLE, *Gandhi: Soldier of Nonviolence,* rev. ed. (1982); and GERALD GOLD, *Gandhi: A Pictorial Biography* (1983).

ROBERT A. HUTTENBACK, *Gandhi in South Africa: British Imperialism and the Indian Question, 1860–1914* (1971), is a study of the Indian community's struggle in South Africa; a study of Gandhi's role in Indian politics and the nationalist movement is presented in JUDITH M. BROWN, *Gandhi's Rise to Power: Indian Politics 1915–1922* (1972), and *Gandhi and Civil Disobedience: The Mahatma in Indian Politics, 1928–34* (1977). SUSANNE H. RUDOLPH and LLOYD I. RUDOLPH, *Gandhi: The Traditional Roots of Charisma* (1983), which discusses Gandhi's remaining influence, was originally published as the second part of the authors' *The Modernity of Tradition: Political Development in India* (1967). FRANCIS G. HUTCHINS, *India's Revolution: Gandhi and the Quit India Movement* (1973), is an interpretive study. GENE SHARP, *Gandhi as a Political Strategist* (1979), is a study of the relation of pacifist principles to political techniques; and JAI CHAND DEV SETHI, *Gandhi Today* (1978), includes an analysis of Gandhian economics.

Among the books containing reminiscences of Gandhi, the more important are: MILLIE G. POLAK, *Mr. Gandhi: The Man* (1931); JAWAHARLAL NEHRU, *An Autobiography* (1936, reissued 1980); SARVEPALLI RADHAKRISHNAN (ed.), *Mahatma Gandhi: Essays and Reflections of His Life and Work,* 2nd ed. (1949, reissued 1977); CHANDRASHANKER SHUKLA (ed.), *Incidents of Gandhiji's Life* (1949); NIRMAL KUMAR BOSE, *My Days with Gandhi* (1953, reissued 1974); ELI S. JONES, *Mahatma Gandhi: An Interpretation* (1948); and VINCENT SHEEAN, *Lead, Kindly Light* (1949). JAMES D. HUNT, *Gandhi in London* (1978), documents his five visits, with little-known details of those in 1906 and 1909. WILLIAM L. SHIRER, *Gandhi: A Memoir* (1979, reprinted 1982), is based on the author's work as a journalist in India in the 1930s.

Among the books highly critical of Gandhi are BHIMRAO R. AMBEDKAR, *What Congress and Gandhi Have Done to the Untouchables* (1945, reissued 1977); CHETTUR SANKARAN NAIR, *Gandhi and Anarchy* (1922); and INDULAL K. YAJNIK, *Gandhi As I Know Him,* rev. ed. (1943). MARTIN B. GREEN, *The Challenge of the Mahatmas* (1978), and *Tolstoy and Gandhi: Men of Peace* (1983), are the first and the last books of the author's trilogy on great leaders and their influence. RAGHAVAN N. IYER, *The Moral and Political Thought of Mahatma Gandhi* (1973, reprinted 1978), compares his concepts to those of Western thinkers. ARNE NAESS, *Gandhi and the Nuclear Age* (1965), and *Gandhi and Group Conflict* (1974), explore basic principles and assumptions of Gandhi's philosophical system. GLYN RICHARDS, *The Philosophy of Gandhi* (1982), explores the relation of his ideas to Hindu metaphysics and to contemporary philosophy. VED MEHTA, *Mahatma Gandhi and His Apostles* (1977), examines the spread of Gandhi's ideas.

There are numerous anthologies of Gandhi's writings. *Selected Writings of Mahatma Gandhi* (1951, reissued 1971), ed. by RONALD DUNCAN; and *All Men Are Brothers* (1959, reissued 1980), ed. by KRISHNA KRIPALANI, are judicious selections for the general reader. *The Words of Gandhi* (1982) is an illustrated selection of quotations, collected and edited by RICHARD ATTENBOROUGH.

Garden and Landscape Design

The vegetated landscape that covered most of the Earth's continents before men began to build still surrounds and penetrates even their largest metropolises. All human efforts to design gardens and to preserve and develop green open space in and around cities are efforts to maintain contact with the original pastoral, rural landscape. Gardens and designed landscapes, by filling the open areas in cities, create a kind of continuity in space between structural urban landscapes and the open rural landscapes beyond. Moreover, gardens and designed landscapes have a special kind of continuity in time. Buildings, paintings, and sculpture may survive longer than specific plants, but the constant cyclical growth and change in plants provide a continuous time dimension that static structures and sculpture can never achieve.

This article discusses the functional aspects of landscaping, the aesthetic and physical components of design, the various kinds of private and public design, and the role and development of gardening in human history.

The article is divided into the following sections:

Functions and concerns of garden and landscape design

ASPECTS OF LANDSCAPE ARCHITECTURE

Garden and landscape design is a substantial part but by no means all of the work of the profession of landscape architecture. Defined as "the art of arranging land and the objects upon it for human use and enjoyment," landscape architecture includes also site planning, land planning, master planning, urban design, and environmental planning. Site planning involves plans for specific developments in which precise arrangements of buildings, roadways, utilities, landscape elements, topography, water features, and vegetation are shown. Land planning is for larger scale developments involving subdivision into several or many parcels, including analyses of land and landscape, feasibility studies for economic, social, political, technical, and ecological constraints, and detailed site plans as needed. Master planning is for land use, conservation, and development at still larger scales, involving comprehensive areas or units of landscape topography or comprehensive systems such as open space, park-recreation, water and drainage, transportation, or utilities. Urban design is the planning and designing of the open-space components of urbanized areas; it involves working with architects on the building patterns, engineers on the traffic and utility patterns, graphic and industrial designers on street furniture, signs, and lighting, planners on overall land use and circulation, economists on economic feasibility, and sociologists on social feasibility, needs, and desires. Environmental planning is for natural or urbanized regions or substantial areas within them, in which the impact of development upon land and natural systems, their capacity to carry and sustain development, or their needs for preservation and conservation are analyzed exhaustively and developed as constraints upon urban design and master, land, and site planning. Within this framework of comprehensive survey, study, analysis, planning, and design of the continuous environment, garden and landscape design represents the final, detailed, precise, intensive refinement and implementation of all previous plans.

Various scales of design

Ideally all of these planning and design phases follow one another closely in a continuous sequential process, but this rarely happens. Various levels of planning and design are performed by different people at different times; often the more comprehensive phases are not performed at all or are performed in an oversimplified manner. The wise gardener or landscape architect, therefore, always begins with a careful analysis of conditions surrounding his project.

Function of garden and landscape design

Garden and landscape design deals with the treatment of land areas not covered by buildings, when those areas are considered important to visual experience, with or without utilitarian function. Typically, these land areas are of four types: those closely related to single buildings, such as front yards, side yards, and backyards, or more extensive grounds; those around and between groups of buildings such as campuses, civic and cultural centres, commercial and industrial complexes; those bordering and paralleling transportation and utility corridors such as parkways, freeways, waterways, power easements; and park-recreation open-space areas and systems. These areas may be of any size, from small urban courtyards and suburban gardens to many thousands of acres of regional, state, or national parks. Although usually conceived as vegetated green spaces on natural ground, they can include also playgrounds, urban plazas, covered malls, roof gardens, and decks, which may be almost entirely formed by construction and paving.

Garden and landscape design, therefore, works with a wide range of natural and processed materials capable of holding up well in the specific local climatic conditions of the site. These materials include earth, rock, water, and plants, either existing on the site or brought in; and construction materials such as concrete, stone, brick, wood, tile, metal, and glass.

ART, SCIENCE, AND NATURE

Garden and landscape design is uniquely concerned with direct relations among art, science, and nature. It operates exactly at the frontier between man and nature, developing transitional connecting zones between the outside limits of buildings and engineering structures and the natural forms and processes that surround them. This is true for large houses and gardens in the country, for regional parks at the edges of cities, for urban and suburban gardens, for urban plazas and roof decks; it is true wherever soil exists to be treated, wherever it may be brought in to fill containers, wherever open spaces are exposed to the weather.

If garden and landscape design is concerned with the relations between mankind and nature, it is largely determined by one or the other of the conflicting philosophies about how human beings do or should relate to nature. People know that they are biologically and physiologically the products of natural evolution; yet their great technological accomplishments lead them to feel that they are above, beyond, or outside nature, that they have conquered and dominated the wilderness and have it now within their power to remake the world. Every work of garden and landscape design reflects one or the other of these conflicting attitudes. The Japanese garden, for example, is inspired by the notion of mankind as a part of nature; the Renaissance garden, by the idea that humans are nature's masters. Garden and landscape design thus reveals much about a culture and a period. One result of the impact of the late 20th-century environmental movement on design in the West may be the emergence of design values that seek the integration of the human and the natural rather than their separation.

Garden and landscape design is an art insofar as it creates for people experiences that uplift their spirits, expand their vision, and invigorate their lives. It is a science insofar as it develops precise knowledge of its processes and materials. And it is directly related to and expressive of nature insofar as it incorporates natural materials and scenes. When the preservation of natural landscape is primary, as in a regional park, art and science manifest themselves in the skill and sensitivity with which necessary facilities and changes are related to the natural landscape. At the other extreme, in an urban plaza, trees in boxes or openings in the paving may be the only natural elements; art and science then are manifested in the design and construction of the total plaza, including its display of trees as symbols of nature, as pleasing forms, and as sources of shade.

Art, science, and nature become most intimately interlocked in certain aspects of horticulture expressed in designed gardens and landscapes: in improved varieties of herbaceous and woody plants; in the cultural practices that stimulate their maximum contributions to the scene; and in the techniques and skills for directing and reshaping the forms of plants—in a range from trimmed hedges and topiary (careful sculptural cutting), through espaliered (trained to grow flat against a wall or trellis) and pollarded (cut back to the trunk to promote a dense head of foliage) trees, to ultimate refinements such as the Japanese practice of removing individual needles from pine trees.

No doubt much art recognizes, expresses, or symbolizes nature in some way. Only the arts of garden and landscape, however, produce works in which nature participates directly with more processed forms and materials.

Functional and technical aspects

As in most other arts, garden and landscape design must solve not only aesthetic but also technical and functional problems. Gardens are for horticulture as well as for viewing, parks are for active recreation as well as for passive relaxation. The surface of the earth must be covered to prevent erosion, dust in summer, and mud in winter. Water persists in running downhill, and even light garden structures must have adequate footings.

This does not mean, however, that there is an inherent or inevitable conflict between utility and beauty. Such conflicts usually develop either because the designer tries to carry out an aesthetic concept that ignores the technical and functional requirements of the problem or because the program is so demanding technically, functionally, or economically that it eliminates aesthetic considerations from the design process. In most garden and landscape design situations it is necessary first to evaluate the technical conditions and functional demands and then to derive from them design concepts that resolve them.

Design

AESTHETIC COMPONENTS

Elements. The traditional elements of design are space; mass; line, or outline; colour; light and shade; texture; scent; and time, as related to climate, season, and growth factors.

Space. Space is air or atmospheric volume defined by physical elements and man's visual imagination. Space has floors: earth, rock, grass, low planting, concrete, asphalt, stone, brick, wood, carpet, tile, linoleum. It has sides or walls: topography, rock, vegetation, vertical structures. And it has ceilings: treetops, structural coverage, or the sky. The most easily understood spaces are the rooms, terraces, patios, and gardens of private residences. A room is defined precisely and unavoidably by floor, walls, and ceiling, particularly with doors closed and windows draped. Beyond these rooms there are the streets, squares, plazas, parks, and public buildings of the city. An urban plaza surrounded by major buildings likewise has positive floors and walls, with sky for ceiling. The fields, meadows, orchards, groves, forests, plains, lakes, river and stream valleys, hills, and mountains of the wider landscape have less precise and regular enclosures. Patios may have fences and walls, gardens hedges and trellises, but in parks there are loose spaces of many soft sizes and forms, defined by trees, ground forms, and shrubbery masses. And in the open landscape there are many different apparent space scales, from the intimate, small farming scenes of New England, Portugal, or Japan to the almost limitless panoramas of the Great Plains, Southwest deserts, or Rocky Mountains. In all of these, space is defined by ground or floor surfaces below, obstacles that block vision horizontally or terminate it at the horizon, and sky overhead. Man's sense of space in all of them results partly from what he actually sees and partly from his imaginative extension, interpretation, and structuring of what he actually sees. Thus, a sand dune, a rock, and a cactus may become a "room" in the desert, while the entire Yosemite Valley is a great room housing thousands of people.

Mass. Mass is the opposite of space. They define each other and depend upon each other for visual existence. Mass may be topographical earth forms, rock outcrops and boulders, trees and shrub groups, buildings, and water forms—streams, lakes, or waterfalls. These are masses in the larger landscape, even though they also incorporate spaces within themselves. Trees, shrubs, and buildings have multiple spaces within them, even though they read as masses from outside. Water forms contain spaces for divers and aquatic life, but of a different density.

Line. Line in the landscape may be the sharp edge of paving, structure, or rock; the boundary between two different surface materials, as grass and ivy; the edge of a shadow; or the silhouette outline of any three-dimensional form, such as a rock, plant, or building. Whatever its source, a line in the landscape plays an important role in the way man sees, interprets, and relates to the scene. A line may lead the eye into the distance, around a corner and out of the scene, or around the scene and back again, holding the viewer within it. It is similar to the role of lines in a painting, holding the viewer within or leading him out of the composition. In a landscape, however, the function of lines is vastly more complicated and difficult to predict. The pattern—that is, the form created by lines—is three-dimensional in any given scene that is viewed. It is four-dimensional in that a spectator continues to move through the landscape over periods of time. The pattern changes throughout each day because of the changing light and shade patterns produced by the movement of the earth around the sun. And the pattern is never exactly the same on one day as on any previous day, because of changes in the weather, the seasons, and the elements of the landscape. Buildings, topography, and rocks may be maintained almost the same for substantial periods of time; but vegetation changes constantly, with both seasonal adjustments and annual growth. That is one reason why landscapes without vegetation seem static, lifeless, and monotonous.

Line in painting and landscape compared

Colour. Colour gives physical landscapes that final dimension of real life, definition, and interest. Spring blossoms and fresh green leaves, after the cold barrenness of winter, herald a new season of vitality and fun. After the deep and stable green of summer, fall colours mark a last resurgence of liveliness before the winter barrenness sets in again. The apparent sizes and forms of landscape spaces change with each such seasonal change: bright colours advance, dull colours recede, changing apparent distances.

Flower garden at Mount Vernon, Virginia, showing the traditional elements of garden and landscape design.

By courtesy of the Mount Vernon Ladies' Association

Structural colours, too, affect the apparent sizes and forms of landscape spaces. Most obvious is the negative effect of bright billboards upon quiet landscapes. To most people billboards seem destructive and arbitrary intrusions; they do not grow out of the scene but are forced onto it. Yet man-made forms—even billboards—can be made to appear to be a part of nature to the extent that they are designed to harmonize with the existing scene.

The aim of the garden and landscape designer is to combine the strong, artificial colours of paint and structure with the softer and more subdued grays, greens, browns, and blues of nature as well as with seasonal outbursts of the purest and truest colours in the world.

Colour variations

Colour varies by hue, the actual colour from the colour wheel; by value, the strength of the colours, bright or pale; the tone or grayness, how pure they are or how grayed by admixture with other colours; by the way that light and shade play on them; and by the texture, smooth or rough, of the surface they are on. All of these factors are taken into account by the garden and landscape designer.

Light and shade. Because the sun—and, to a lesser extent, the moon, stars, fire, and artificial lighting—has the property of casting shadows, landscape design, in placing trees, structures, and other elements on the land, must always take into consideration the light and shade resulting from such placement.

Light and shade are not the same in all parts of a country or the world. Light is welcome in cool, gray, northern climates, shade in hot, bright, desert or tropical regions. In the clear air of unspoiled deserts, man sees so far that he loses all sense of size, scale, and distance; in the foggy humidity of the western coasts of Europe and North America distances seen and objects perceived change from day to day, sometimes from hour to hour, so that one lives with a continuing sense of mystery and variety. Landscape design must, ideally, remain sensitive to and work carefully with the light and shade relations that are most desirable in each different region or subregion.

Texture. Texture—the smoothness or roughness of surfaces—is another element of landscape design. It is perceived primarily by touch, although through vision one approximates the textures of different surfaces and imagines how they would feel. The surface texture of the earth may vary from fine sand or silt to coarse clods, gravel, or boulders. The texture of plant coverage may vary from fine bent grasses through coarser meadow grasses to brush, ivy, or cactus. Wall surfaces range from the smoothness of glass and plaster to the roughness of brick, stone, or rough-sawn lumber.

Tactile textures must be experienced intimately. Visual textures may be experienced at any distance. Farther away, larger elements participate in texture effects; at medium distances the foliage of trees and the size of rocks create textural qualities; from an airplane or hilltop the size and arrangement of buildings, topographic forms, masses of vegetation, or water create textural effects.

Scent. Scent is a delicate and subtle element in landscape experience, often lost to 20th-century man because of widespread pollution of the air with foul-smelling exhaust and waste gases. The fragrance of flower and fruit is one of the traditional delights of garden and park, still attainable through sensitive selection and arrangement of plants.

Time, climate, and season. Unlike the static continuity of architectural and urban monuments, garden and landscape spaces are dependent on maintenance, which determines whether the form envisioned by the original designers will endure or change over decades or centuries. The Saihō-ji garden and many others in Japan continue today in much the same form as they began because of continuing maintenance. On the other hand, the inadequate maintenance of the Renaissance gardens—designed as geometric architectural abstractions, to which plant forms were made to conform by clipping and trimming—has allowed many of the larger plants to resume their natural forms. The results, however pleasant, are not what the designers envisioned. Instead of hundreds of years, the typical suburban garden in the United States has a predictable life of about five years. Ownership or tenancy tends to change in that cycle, with unpredictable results in the garden.

Maintenance of gardens

Time and climate are closely interrelated in their effects on garden and landscape spaces. Because relations between temperature and moisture, light and shade, change daily throughout the year, every region and locality on earth—in fact, every building site—has its own climate, unlike any other. Therefore, garden and landscape design for every region, locality, and site may be expected to be different. Nevertheless, climates can be generalized in certain broad categories that are similar, though not identical, over large areas of the earth.

Climatic factors having major impact on garden and landscape design are temperature range (hot to cold), precipitation range (high to low, rain to snow), their combinations (hot, wet summers; hot, dry summers; cold, wet winters; cold, dry winters; and so on), growing season (year around in the tropics, a few weeks in the Arctic), atmospheric humidity (clear air or clouds, fog, mist) and its effects on visibility and light-shade patterns, air movement (winds, breezes) and its effect on the other factors (cooling in hot weather, chilling in cold, moving clouds and fog).

The combination of all of these factors affects how one sees the landscape (bright, clear desert distances; soft, mysterious, changeable foggy landscapes), what one expects

Grand east–west axis of the gardens at Versailles, France, showing the basic principles of garden and landscape design. Designed by André Le Nôtre (1613–1700).
Rapho/Photo Researchers

from design (shade from the sun, protection from the wind, shelter from rain and snow), and how one designs gardens and landscapes. The patios, cloisters, and oases of Mediterranean and Middle East regions, the romantic naturalistic parks of Europe, China, and eastern North America, the esoteric garden abstractions of Japan—all of these different approaches to design were inspired partly by the particular qualities of the landscape climate in which they developed.

Time and natural light are, of course, intimately interlocked in the daily cycle of night and day, in the seasonal cycles (light is different in summer and winter because of differences in temperature and humidity), and in the annual cycles (long days in summer, short days in winter).

Time, climate, and season are all reflected in garden and landscape in the growth and change of plants. In the tropics, growth is constant and taken for granted, a problem of control. As growing seasons become shorter in the north and south or at higher elevations, they become more precious. In far northern and southern latitudes the short summer is a period of rejoicing and outdoor activity. A tree that will mature in five to ten years in Southern California or Florida will require 30 to 40 years in the North Central states. Spring blossoms, fall colour, the change from summer green to winter's exposed branch structures, all of these mark the seasonal changes clearly and strongly in the Eastern United States, in Europe, and in other temperate areas.

Principles. The basic principles of design deal with the arrangement or organization of the elements, as expressed in specific materials, on a specific site. The site—a piece of land, with surface form and internal content—may in itself require reshaping. On it will be placed—or may already exist—buildings, roads, minor structures, trees, shrubs, ground-cover planting, water elements, rocks.

The elements of design are contained in these individual components and in specific relations that may develop among them on a particular site. The principles of design—which deal with overall relations—are unity and variety, rhythm and balance, accent and contrast, scale and proportion, and composite three-dimensional spatial form.

Unity and variety. Unity and variety are derived from the number of elements, or kinds of material, within a given visual area and from the way they are combined. A brick building or a rose garden is unified by concentration on one material. The difficulty of achieving a sense of unity increases as the number of elements, or kinds of material, increase. A building of six materials or a garden of 30 kinds of plants, for example, will have more variety but unity can be achieved only by careful organization and arrangement. At a certain point, which varies with the situation and the skill of the designer, it becomes impossible to establish unity. Variety then dominates.

Rhythm and balance. Rhythm and balance result from the three-dimensional arrangement of elements and materials on the site. Rhythm is a sequence or repetition of similar elements—as a double row of trees. It tends to emphasize direction and movement, as along an allée toward a viewpoint or terminus. Balance is the sense one gets, looking in any direction, that the elements to one's left balance those to one's right and the feeling one has that the views one has just experienced are in equilibrium with what one sees now. The most obvious examples of balance are the symmetrical axial Renaissance gardens of Europe, but these are not the only or even the most interesting ways to achieve balance.

Difference between accent and contrast

Accent and contrast. Accent and contrast enliven arrangements that may be so balanced, orderly, and harmonious as to be dull. An accent is an element that differs from everything around it, as silver-gray foliage against dark-green conifers, but is limited in quantity in relation to surrounding elements. Contrast is stronger: two different elements may be juxtaposed in almost equal quantity to emphasize the special qualities of each. Well-known examples are the formal palace in the informal park, the green park in the densely built-up city. Accent and contrast are more difficult to handle successfully than straightforward, simple, harmonious design. An example of the failure to handle it successfully is the common practice of lining a street with alternate specimens of two quite different trees, as pines and cherries, which merely cancel each other out.

Scale and proportion. Scale refers to the apparent (not the actual) size of a landscape space or of the elements within it. Proportion is the determined relations among the sizes of all the parts within an element and of all the elements within a space. Thus, the proportionate sizes of the legs, arms, and back of a garden bench, for example, determine the scale of the seat. And the overall size of the seat, in proportionate relation to walk width, arbor height, lawn area, tree size, and so on, helps to determine the scale of the garden.

Composite three-dimensional spatial form. Composite three-dimensional spatial form results from the delineation of a block of air by physical elements, which enclose and frame the space and establish its relations with neighbouring spaces, distant views, and so on. A patio with paved floor and walled enclosure (with perhaps a grilled outlook) and sheltered by trees or pergola structures (arbor or trellis) is an obvious example of this form.

The design process. The design process has been called in the past modes of composition and style or period selection. In the first quarter of the 20th century, the arts, including architectural, garden, and landscape design, were dominated by traditional, eclectic, preconceived systems of form and approach called the Beaux Arts system, after the famous school in Paris. In essence, these systems told designers what to design and where. Their only choice and their only skill lay in how to adapt preconceived systems—such as formal and informal gardens—to the particular problem at hand. Innovation consisted of timid new relationships among traditional elements.

The modern revolt in design

In the first quarter of the century there also occurred what is called the modern revolt. Beginning in painting and sculpture, it soon swept through architecture and reached garden and landscape design toward the end of the quarter in Europe, reaching the United States in about 1935.

The essence of the modern revolt was the rejection of preconceived or traditional styles, periods, rules, regulations, or systems governing design. In place of these, systems and processes developed for analyzing problems and situations in their own terms and in terms of the modern resources available for solving them. Basic to the new theories was the idea that designed forms should arise from and express each specific situation and the contemporary, industrial culture around it. By the 1970s all fields of design seemed to be dominated by these theories; but, although submerged, traditional Beaux Arts design continues to surface regularly in strange new combinations with modern forms. A form of this eclecticism emerged in the early 1970s, when architects once again designed symmetrical monumental buildings with little functional or structural expression, and traditional formal–informal concepts in garden and landscape design began to reappear.

PHYSICAL COMPONENTS

Natural. Natural integrants of garden and landscape design include earth, rock, water, and plants.

Earth. As a base for design, earth is the floor of landscape spaces, the root medium in which half of every plant lives, the foundation for structures, the vehicle for surface and subsurface drainage of excess water, and a sculptural material in its own right.

As a floor, earth can be seen as an abstract surface. If apparently level, with just enough slope for drainage, it is ready to be covered with paving, grass, ground cover, or other planting, which is necessary to prevent dust in dry weather and mud in wet weather; if sloping or irregular, earthwork may be necessary to conform to new construction or to the design plan, to provide adequate drainage, or in order to relate properly to neighbouring topography and views.

Understanding earth as soil

As a root medium for plants, earth must be understood as soil. One must know the type and depth of soil before planning a garden or landscape. Soil occurs in layers: topsoil, in which there is a high percentage of organic humus and micro-organisms; subsoil, which is more sterile as it gets deeper; and bedrock, which is not yet broken up. There are many variations in these layers. In the mountains there may be only a few inches of soil over rock; in old valleys the soil may be hundreds of feet deep. Most plants require one to six feet of topsoil, with good drainage, but there are plants that will grow in rock, sand, sterile soil, boggy land, shallow water, or open water. If the soil is not adequate for the planting desired or if the form of the earth is to be changed, then new soil conditions must be created.

As a foundation for structures, earth must be dry and firm. Although structures can be built in almost any soil, they become more and more expensive as the earth becomes less dry and firm. Desirable foundation conditions, the exact opposite of the loose, moist soil that is best for most plants, create many technical problems in the relations between structures and plant areas.

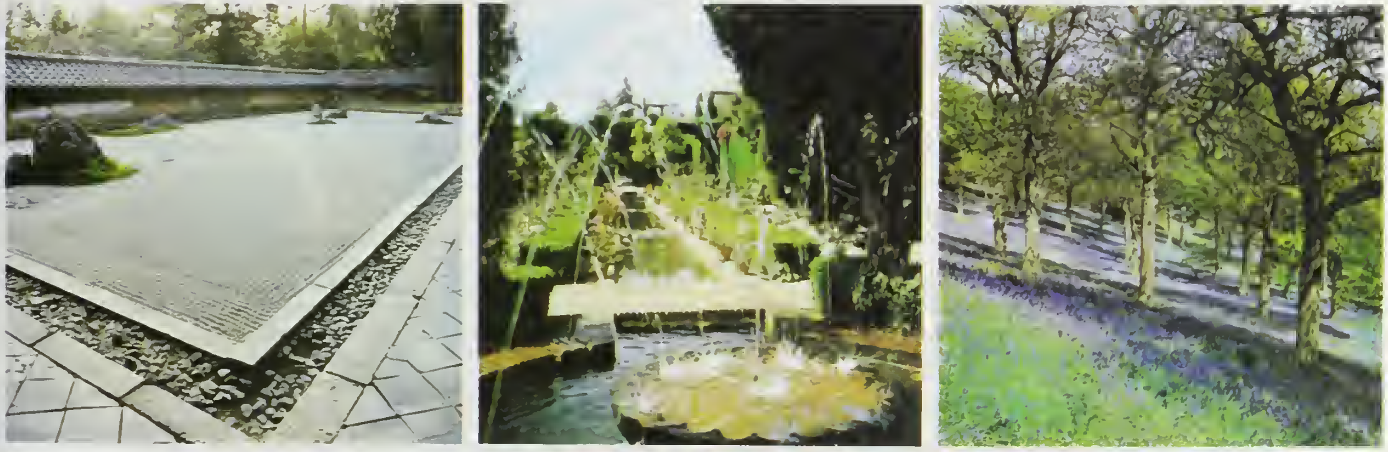
As a drainage vehicle, earth absorbs a high percentage of the water that falls on its surface. This absorbed water may be stored below ground, or it may move horizontally through sloping soil patterns. Surface water that is not absorbed, either because the soil is saturated or because the slope of the ground makes it run off too fast, must drain away on the surface. This creates many technical problems, especially if the surface is not covered to prevent erosion or if a great deal of land is covered by roofed structures or paved surfaces, which increase the amount of water running off because none is absorbed.

As a sculptural material, earth can be contoured to conform with functional and maintenance demands. Rolling natural hills and golf-course earth forms demonstrate the potential. Slopes must not be too steep for planting to hold, unless they are structurally retained.

Rock. Rock is a major factor in some regions, minor in others, nonexistent in some. It varies in size from sand through pebbles, cobbles, boulders, and fixed outcrops to solid-rock mountains. It varies in form from square or jagged, newly cut or broken to rounded forms produced by the action of water. It varies also in colour and in texture. It can be used as a ground cover, dry or in cement; in vertical structures with various degrees of cutting and finishing; to simulate natural rock formations; and in sculptural groupings that emphasize the natural form of the rocks, as the Japanese do so well.

Water. Water is essential to all gardens and landscapes,

By courtesy of (left) the International Society for Educational Information, Tokyo; (right) the National Trust; photographs, (centre) Ewing Galloway; (right) Andy Williams



The natural and physical integrants of garden and landscape design.
(Left) Rocks, raked sand, paving, and buildings—Garden of the Ryōan-ji, 15th century, Kyōto, Japan. (Centre) Water, plants, and structural pools—Gardens of the Generálife, Granada, Spain. (Right) Sloped earth, trees, and colourful ground cover—Bluebell Wood, Winkworth Arboretum, Surrey, England.

Garden of the Kinkaku-ji (Golden Pavilion), showing the use of a shelter structure, the pavilion, as the main focal point of a landscape design, 15th century, Kyōto, Japan.

By courtesy of the Consulate General of Japan, Chicago

even in the desert, although amounts vary considerably. As a design element, water contributes coolness, moisture, sparkle, lightness, depth, serenity, the possibility of aquatic plants and animals, and recreation. It may run through natural stream and river channels and collect in natural ponds, lakes, and seas, or it may be kept in structural channels and pools, recirculated to avoid waste. In some soil conditions it may be necessary to seal thc bottoms of natural-appearing streams and ponds to hold the water. Under natural conditions water runs downhill through naturally formed channels and collects in low spots or bowls of firm form. Water also may be pumped uphill or thrown into the air in jets and fountains. Water is used increasingly throughout countries with technically complex artificial irrigation systems, which become distinctive landscape elements.

Aesthetic qualities of water

Plants. Plants are considered the primary material of gardens and landscapes, by contrast with the concentration of structures in cities. They may be grouped and organized for design purposes in several ways: by size (trees, shrubs, low plants, grass); by form (vertical, horizontal, round, irregular); by texture (size, shape, and arrangement of foliage and structure); by colour (flowers, fruit, foliage, structure); by rate of growth (fast, medium, slow); by seasonal effect (spring, summer, fall, winter); by fragrance (flowers, fruit, foliage); by environmental requirements (soil, drainage, sun–shade, temperature range, pests and diseases, pruning needs). All of these properties affect the selection, arrangement, and maintenance of plants in designed landscapes.

Structural. Structural integrants of garden and landscape include structures closely related to the earth, enclosure structures, shelter structures, engineering structures, and special buildings.

Earth-related structures. Earth-related structures include paving (walks, roads, terraces, patios) and change-of-level structures (retaining walls, steps, ramps, bridges), which must be made of materials that will resist decay, such as brick, stone, concrete, asphalt. These structures provide the connections for movement and circulation and the areas for intensive gathering, social use, or active recreation. They embody a complex technology.

Enclosure structures. Enclosure structures, such as walls and fences, are designed to control vision or movement or both. They may be of various heights, three to ten feet (one to three metres) or more, and of many materials: brick, stone, or concrete masonry; wood; metal; sheet materials such as glass, plastic, asbestos, pressed boards. Because they are at eye level and extend and connect buildings, they are very important in intimate visual design.

Shelter structures. Shelter structures, designed to protect from sun, rain, or wind, may incorporate enclosure elements at the sides with overhead elements, which may be open framework pergolas or arbors carrying vines or solid opaque or translucent roofs. Among such structures are gazebos, pavilions, garden temples, summer houses, hermit huts, follies, ruins, and grottoes.

Engineering structures. Engineering structures tend to appear unexpectedly, because incongruously, in gardens and landscapes. They are usually mechanical or electrical transformers, vents, valves, siphons, drains, culverts, headwalls, dams, and many other nameless and mysterious forms. In a good landscape design, these structures are integrated into the overall plan.

Special buildings. Special buildings include many that are nearly as complete as the fully enclosed buildings that are the province of the architect: greenhouses, conservatories, orangeries, tool sheds, dovecotes, icehouses, root houses, bathhouses, playhouses, and many more. These are usually auxiliary in relation to the main house or building but relate to them in character and detail.

Shelter structures, engineering structures, and special buildings are important in garden and landscape because they introduce precise geometric forms intermediate in scale between main buildings and landscape. If scattered or designed indiscriminately, they can destroy a pleasant landscape; carefully designed, they can be so grouped or arranged as to create rhythmic connections and patterns within the overall architectural-landscape design concept.

Importance of structures and buildings in design

Sculptural components and outdoor furnishings. Sculptural components of garden and landscape have traditionally been predictable forms and types: figurative sculpture, decorative urns and plaques, fountains, sundials, birdbaths, cisterns, and wells. All of these continue to appear as elements of the persistent underground Beaux Arts vocabulary. In contemporary design, however, they are eliminated or take on new forms derived from modern sculpture. The possibility now exists for the production of gardens and landscapes so completely sculptured that one cannot tell where design stops and sculpture begins.

Outdoor furnishings and equipment include all of those fixed and movable elements that tend to appear in garden or landscape after the plans are done and installed and therefore without benefit of design control. In the garden they are seats, tables, barbecues, umbrellas, plant containers, supports, and guards, as well as lights and light systems. In the public landscape they include these garden elements and many more: signs, trash containers, alarm boxes, mailboxes, newsstands, kiosks, service elements, telephone stands. If all of these elements are not predicted insofar as possible in the original design plans, incorporated in them, and carefully controlled thereafter, they can destroy carefully planned landscapes. One red oil drum for trash can dominate the visual experience of a large, pastoral picnic area.

Kinds of design

The landscape is everything an observer, whether still or in motion, can see. The landscape as a work of individual art is any garden or space designed, developed, and maintained for the private experience of an individual or family, a space not accessible to others either physically or visually. The landscape as a work of collective art is everything beyond this private range: everything seen beyond the confines of private gardens or estates, all borrowed landscapes, all streetscapes, all city, metropolitan, and regional landscapes, and their accumulation in national, continental, and world landscapes. This collective art may be good or bad depending on whether it results from the accidental accumulation of individual and conflicting efforts or from controlled and planned efforts.

PRIVATE OR RESIDENTIAL DESIGN

The history of landscape design is largely the history of landscape as a work of private, individual art. Plazas (structural public open spaces not dominated by foliage), throughout classical, medieval, and Renaissance history, were the concessions of the ruling class to the need for public meeting places; but it was not until Central Park was developed in New York City in the mid-19th century that this need reached the level of designed public green spaces. During most of its history, landscape design was of three kinds: private utilitarian farms and gardens; private gardens in which the enhancement of the quality of living was paramount; and private gardens designed to express the power and benevolence of the ruling or upper classes. The expansion in scale of private gardens beyond the needs of private living led inexorably, first, to the dedication of such spaces to public use and then to the development of public gardens and parks designed for public use.

Kinds of private gardens

The private garden, however, has remained the centre for private fantasy and a means of escape from the grinding and difficult world of reality. The most important aspect of the private garden is its seclusion: from the physical world, by means of distance and enclosure; from the social world, by separation and exclusion. Space and greenery are also important. The space may be very small, perhaps a tiny courtyard, and greenery limited to one or two plants, but these make possible that private world of fantasy that may make the difference between sanity and lunacy. The 20th-century mass migration to the suburbs is the latest expression of this need.

Generally, the private garden occupies a space somewhere between 20 feet (six metres) square and one quarter of an acre (100 feet square). The forms of private gardens range from the formalism of pure geometry or the artistic representation of natural processes through the variations of standard gardening techniques and the informalism of letting nature take its course to various manifestations of literary, poetic, historic, and subjective concepts.

When housing moves from single-family, detached buildings on private lots to higher density variations—duplexes, semidetached villas, town houses, clusters, condominiums, low- and high-rise apartments—new relationships develop. As population density increases, private design shrinks and public design increases. Somewhere between the extremes of the single-family dwelling with minimum public space and the high-rise apartment with minimum private space, there is an optimum relationship in which real needs can be expressed. Perhaps the best potential lies in town-house, cluster-house, and condominium developments, in which there is a flexible relationship between public and private elements.

PUBLIC DESIGN

Because of fixation on the notion that the original resource of land and landscape, continuous from sea to shining sea, is best organized for private or public use by gridiron subdivision into innumerable separate parcels, public landscape design begins at the level of single buildings on single lots, with front yards and backyards. The buildings may be government offices, quasi-public companies, or private corporations, but all tend to be designed in terms of public and private spaces, as though they were private residences for the groups involved.

Campus design

Campus design begins when publicly accessible buildings grow into complexes of two or more, for religious, commercial, industrial, governmental, or educational use. Instead of or in addition to simple front-yard and backyard design, there are more complex systems of spaces between buildings, which vary from courtyards and quadrangles of varying forms and dimensions to passageways connecting them in varying widths and degrees of overhead coverage. The open spaces range in character from paved architectural courtyards and cloisters to open playing fields and parklike spaces. Campus design makes possible the richest, most complex, and rewarding range of relationships between architectural and landscape design. Perhaps the best examples, in which the sequential experience of indoor and outdoor space approaches the maximum, are the religious, educational, and civic complexes of Europe, developed before the idea of gridiron subdivision fragmented environmental design. In China and Japan there are many highly refined and sophisticated temple, shrine, palace, and castle complexes. There are also many fine examples in the United States of similar institutions that have transcended or resisted subdivision.

Urban design

In the broader area of urban design, landscape architecture deals with such open-space components as public gardens, parks and playgrounds, plazas, squares, and malls. In these urban spaces, the designer attempts to meet the need for community, for play and recreation, for refreshment and relaxation, for individual withdrawal in a gregarious atmosphere.

Towns, cities, and metropolitan areas may be said to have three basic components: buildings, designed by architects or builders; open spaces, designed by landscape architects or technicians; and circulation-utility corridors—street, highway, railway, and rapid-transit systems—which are usually planned and designed by engineers.

The basic structure of urban areas consists of the open spaces together with corridors comprising a total open-space system, defined by and connecting the buildings. The corridors have usually been considered merely a utilitarian framework, connecting and servicing buildings and quality open spaces, channelling traffic and utilities throughout urban areas, and connecting them with the open country around. Modern urban thinking has begun to go beyond this concept, to see the total open-space-corridor system as the major qualitative structure of the city, which, when viewed in conjunction with overall building design, is seen to establish the city's basic character.

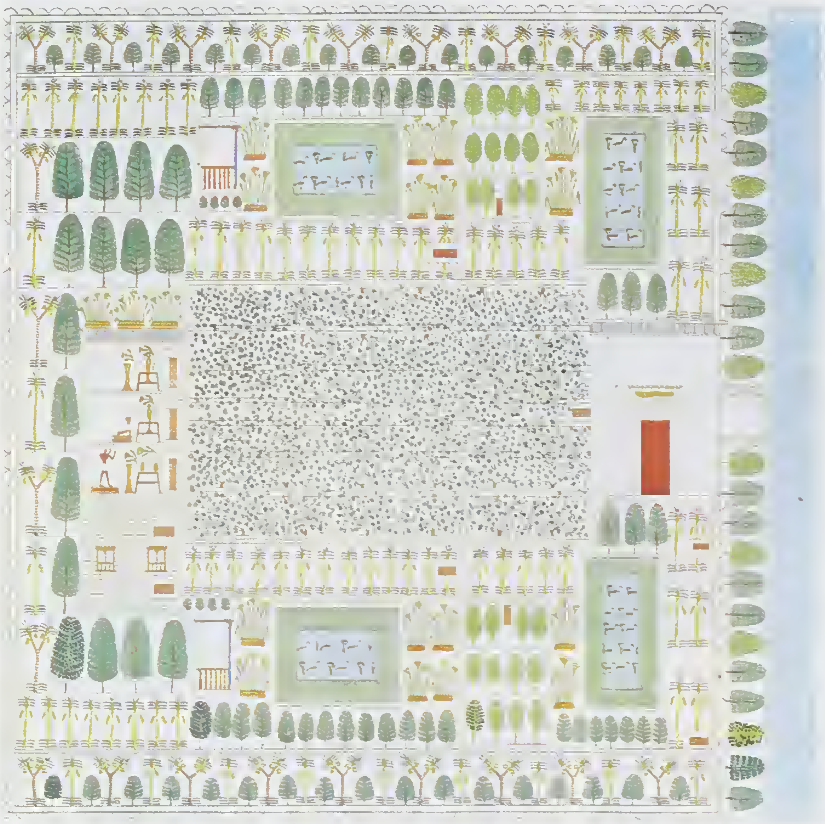
From this point of view the role of landscape architecture, once limited to tasteful planting of corridors designed by engineers, begins to expand. Some urban planners would expand it even further, believing that open-space corridors should be designed throughout primarily as social spaces for people and only secondarily as utilitarian passages for vehicles.

Commemorative sites—cemeteries, historic spots, battlefields—are important because they memorialize and symbolize important events in personal, local, national, or world history. Wherever they occur, these sites or areas are marked with stone or bronze memorials and often dramatized with more elaborate developments. The designs tend to follow traditional and conservative precedents, often impressive but seldom imaginative. It is still difficult to equal Asplund's Forest Crematorium (1940) in Stockholm or the Fosse Ardeatine memorial in Rome. (G.Ec.)

Historical development

WESTERN

Antiquity. *Egyptian.* The earliest surviving detailed garden plan, dating from about 1400 BC, is of a garden belonging to an Egyptian high court official at Thebes. The main



Earliest surviving detailed garden plan, the estate of an Egyptian official, *c.* 1400 BC. From Ippolito Rosellini, *I monumenti dell' Egitto e della Nubia*. In the New York Public Library.

By courtesy of the New York Public Library, Oriental Division, Astor, Lenox and Tilden Foundations

The earliest surviving garden plan

entrance is aligned on a pergola (trellis-bordered) walk of vines leading directly to the dwelling. The rest of the garden is laid out with tree-lined avenues, four rectangular ponds containing waterfowl, and two garden pavilions. Although rigidly symmetrical, the garden is divided into self-contained, walled enclosures, so that the symmetry of the whole could not have been apparent to the viewer. Such a highly developed pattern argues a considerable incubation period, and it is likely that similar enclosed pleasure gardens had been designed as early as 2800 BC.

Assyrian, Babylonian, and Persian. The gardens of Assyria, Babylon, and Persia were of three kinds: large, enclosed game reserves such as the garden of Eden described by the Hebrews in the Old Testament; pleasure gardens, which were essentially places where shade and cool water could be privately enjoyed; and sacred enclosures rising in man-made terraces, planted with trees and shrubs, forming an artificial hill such as the Hanging Garden of Babylon.

Greek and Hellenistic. The urban life of ancient Greece led to houses built around central, private courtyards. Lined with colonnades that gave access to the rooms of the house, the courtyard, or peristyle, was open to the sky and insulated from the street. In the peristyle was a garden consisting of a water supply and potted plants. Much of life, however, was lived in public. The sports grounds, where exercise was taken, became popular gathering places and developed into the original academy and lyceum, which included the exercise ground, seats for spectators, porticoes for bad weather, statues of honored athletes, and groves of shade trees. These public recreation grounds set the type for the later classical Roman villa garden and the 19th-century European public park. A third type of Greek garden was the sacred landscape, such as the Vale of Tempe or the mountain sanctuary of Delphi.

The relatively austere Greek taste was transformed in the Hellenistic Age (*c.* 323 BC–30 BC) by the influence of the East. Luxurious pleasure grounds were made, especially at colonies such as Alexandria and Syracuse. These gardens were conspicuously luxurious in their display of precious materials, and artificial in their use of hydraulic automata.

Roman. Roman gardens derived from the Greek, those in the seaside resorts of Pompeii and Herculaneum (1st century BC) following the Hellenistic pattern. These small, enclosed town gardens were visually extended by landscapes painted on the walls. Throughout the imperial period, the more ambitious villa gardens flourished in many forms on sites carefully chosen for climate and aspect.

The most elaborate was that of Nero's Golden House, which covered over 300 acres (120 hectares) in the middle of Rome and included an artificial lake (where the Colosseum now stands) and a pastoral landscape of plowland, vineyard, pasture, and wood. More influential in later times was the vast garden complex of Hadrian's Villa, of which extensive ruins can still be found near Tivoli.

Middle Ages. The barbarian invasions of the 4th and 5th centuries AD destroyed Roman civilization and with it the gardens of western Europe. The Eastern Empire, centred on Constantinople, retained its hold on Greece and much of Asia Minor for another millennium; and Byzantine gardens persisted in the Hellenistic tradition, laying more emphasis on wonder-provoking apparatus than on aesthetic values. A recurrent feature of these gardens was a tree of gold or silver equipped with birds that flapped their wings and sang and branches that sprayed wine or perfume.

Byzantine gardens

Islāmic. Beginning in the 7th century, the Arabs progressively captured much of western Asia, Egypt, the whole of the North African coast, and Spain; in the process, they spread features of Persian and Byzantine gardens across the Mediterranean as far as the Iberian Peninsula. Most characteristic of these gardens was the use of water—the ultimate luxury to desert dwellers, who appreciated it not only because it allowed plants to grow but also because it cooled the air and gratified the ear with the sound of its movement. It was commonly used in regularly shaped, often rectangular, pools. The water was kept moving by simply designed fountains and fed by narrow canals resembling agricultural irrigation channels. Because water was rarely abundant, the pools were shallow but increased in apparent depth by a blue tile lining.

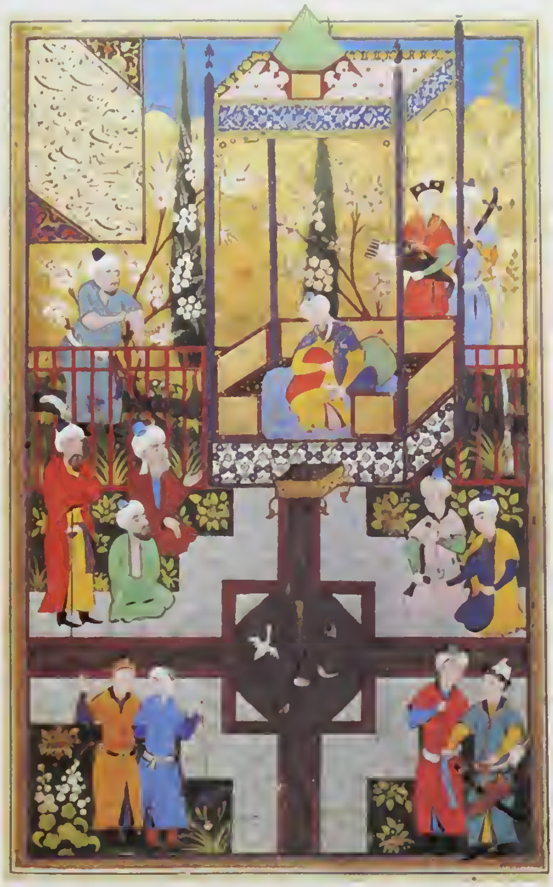
Frescoed wall of fruit trees, palms, and oleanders from the garden room, Villa of Livia, Rome, c. 50 BC. In the Museo Nazionale Romano, Rome.

SCALA—Art Resource

These pools of water graced Islāmic gardens—such as those of the Alhambra in Granada—that resembled the Hellenistic colonnaded courtyard. The gardens provided shade, excluded hot winds, and created the sense of being in a jewelled private world. Water mirroring the sky gave an impression of spaciousness and introduced lightness, brightness, and an air of unreality.

In the Moorish Caliphate of Córdoba in Spain, in the valley of the Guadalquivir, there were said to have been 50,000 villas, all of which probably had such garden courts.

By courtesy of the Metropolitan Museum of Art, New York, Hewitt Fund, 1911



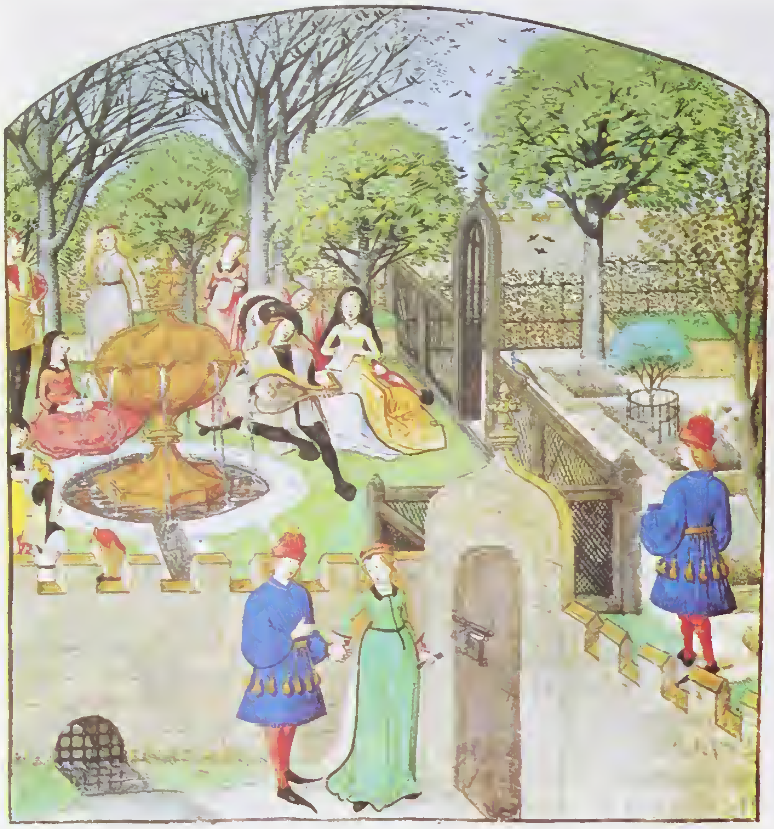
Shallow tiled pool and water channels characteristic of the Islāmic garden. "Sultan of Syria Holding an Audience in His Garden," Persian miniature from the manuscript (1522–23) of the *Būstān* by Sa'dī. In the Metropolitan Museum of Art, New York.

The greatest period of garden making in the Islāmic world was the 14th century. In the vicinity of the conqueror Timur Lenk's capital of Samarkand, the names of 11 royal gardens are recorded, and there were probably others belonging to his nobles. Whereas gardens of the Alhambra type were architecturally conceived within the total plan of a building, some of the more extensive Timurid gardens and their derivatives, the Mughal gardens of India, were pleasances of water, meadow, trees, and flowers, in which buildings took a subordinate place. Although these garden buildings were permanent, their subordinate role and the lightness and luxuriating frivolity of their design mark them as heirs of the casually positioned tents seasonally erected in hunting parks. There were also gardens of strictly architectural design—huge walled enclosures with corner towers, a central palace, regularly disposed avenues, and tanks of water. Deer and pheasants were kept in these gardens, which combined the quality of hunting park and of *hortus conclusus,* or closed garden. Trees were planted sometimes in regular quincuncial patterns (one in the middle and one at each corner of a square or rectangle) but more often freely. In all types of Islāmic gardens, flowers were lavishly used. Their presence was even simulated in garden carpets and in the woven hangings that were used as temporary screens.

Parks made by the Saracen emirs of Sicily

Influential on later Western practice were the parks made by the Saracen emirs of Sicily. The Normans who conquered the Saracens in the 11th century adopted the manner of life of those they had overthrown, and thus the emirs' gardens survived their makers. A large area of the Conca d'Oro, the great natural amphitheatre behind Palermo, was taken up with pleasure grounds—walled enclosures large enough to contain woods and hills, canals, artificial lakes, groves of oranges and lemons, fountains, water stairways, and wild creatures running free.

Western European. In Europe, beyond the limits of the Islāmic conquest, the destruction of civilized society by the barbarian tribes had been nearly complete; but the physical remains of the past shaped the reviving future: the peristyle gardens of Roman villas became the cloisters of Christian basilicas. Security and leisure existed only in the monastic system, which also preserved some of the traditional skills of cultivation. For some time the only type of garden was the cloister, with its well, herbs, pot plants, and shaded walk. Then secular gardens began to appear, but they were usually of limited extent, confined within the fortifications of a castle and often raised well above ground level on a battlemented turret. These Gothic gardens were rectangular, with the traditional division into four parts by paths, the quarters again subdivided according to the amount of ground available and the convenience of cultivation. At the point of principle intersection was a well, which, when elaborated, became the vertical feature of the garden. Seats—often of turf—were constructed in



Medieval walled garden combining a grassy and shaded pleasure area with an herb garden. *Roman de la Rose*, miniature from a 15th-century French manuscript. In the British Museum.

By courtesy of the trustees of the British Museum

the walls. Many flowers were grown, but their season was short; after June and often earlier, the beds were flowerless. More extensive and elaborate gardens were rare.

Italian gardens of the Middle Ages

In 13th-century Italy, through the influence of the Holy Roman emperor Frederick II, who had spent much of his youth in Sicily, the example of the Saracen emirs was felt in Apulia and Naples. The *Triumph of Death*, painted by the Florentine artist Andrea Orcagna (Pisa, Campo Santo), shows a garden of considerably greater extent than the cloister or battlement type. Gardens like this existed also in Lombardy, where the court of Gian Galeazzo Visconti, the founder of the great walled park of Pavia, cultivated the arts of civilized life. In describing the Royal Gardens at Naples, the writer and poet Giovanni Boccaccio speaks of statues disposed regularly around a lawn, interspersed with marble seats. Such a garden suggests that Frederick II's classicizing influence extended into the mid-14th century. Also significant was the garden of Hesdin in Picardy, which became famous throughout France for its automata and water tricks. It was made by a crusader who, having returned to France by way of Palermo in 1270, no doubt incorporated in his garden what he had seen of Saracenic gardens there and in Syria. Hesdin was an exotic creation without parallel in its northerly location for several centuries.

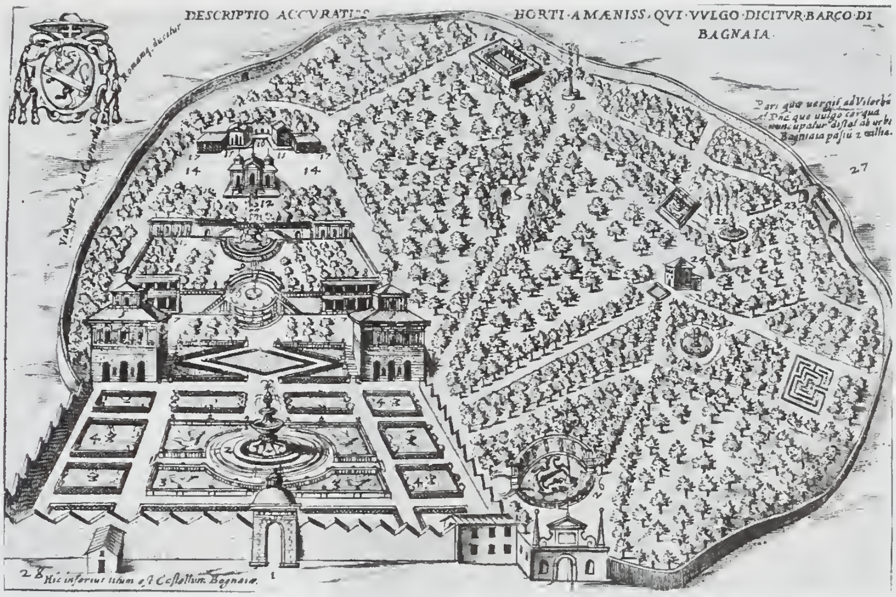
Renaissance to modern: 15th to 20th centuries. *Italian.* The increasing prosperity of western Europe and man's increasing confidence in himself and in his capacity to impose order on the external world was reflected in the gardens of Italy by the mid-15th century. The change began near Florence, where the old medieval enclosures began to open up. The rectangles, which had been dissociated, were now sited one behind the other, thus prolonging the main axis, which was now aligned on the centre of the dwelling. This change inevitably introduced the idea that house and garden were a coherent, complementary whole. And, because villas were increasingly sited for amenity rather than defense, gardens became less enclosed, more susceptible first to visual, then to actual extension.

The unity of house and garden, together with the need for physical adjustment to the sloping sites favoured by classical precedent, threw the planning of the new Renaissance garden into the hands of architects. Most influential was the garden courtyard designed by Donato Bramante at the Vatican to link the Papal palace with the Villa Belvedere; the uneven site and the disparity in bulk of the two buildings was overcome with terraces and stairways. It remained an enclosed garden but one far removed from the earlier cloistral courtyards. The garden of the Belvedere combined the function of an open-air room with that of an outdoor sculpture gallery.

The typical Italian Renaissance garden

The ingredients of the Renaissance garden thus separately established were united in varying proportions. The typical evolved garden of the period was characterized by some openness of aspect, axial development, a tendency to prolongation, unity of concept between house and garden emphasized by a considerable "built" element of stone, lavish employment of statuary (often in the form of fountains), and the proliferation of such classical accents as grottoes, nymphaea (Roman buildings with a fountain, plants, and sculpture), urns, and inscriptions. There is no adequate evidence that this type of garden had an exact equivalent in the classical period, although there is evidence that each of its elements existed.

The variation in style among Italian gardens is considerable and is due not only to the date they were made, the exigencies of the site, and regional variation but also to their social function. The scale of the garden compartments at the back of the Villa Gamberaia at Settignano (1610), for example, is small in contrast with the extensive view over Florence from the front and thus suggests intimate use by members of a small household. The more extensive parterre garden (an ornamental garden with paths between the beds) of the Villa Lante at Bagnaia (begun 1564) is designed neither for solitary enjoyment nor for a crowd but for a select, discerning company—as is the garden of the far more splendid Villa Farnese at Caprarola (completed 1587). The most remarkable mid-16th-century garden, that of the Villa d'Este at Tivoli (1550), is situated on a steep slope of the Sabine hills. The river that plunges down this slope is harnessed to an astonishing variety of fountains, including a "water organ."


DESCRIPTIO ACCVRATIS HORTI AMÆNISS. QVI VVLGO DICITVR BARCO DI BAGNAIA

Plan of the Villa Lante at Bagnaia, Italy, showing the parterre gardens and enclosed park. From Giacomo Lauro, *Antiquae Urbis Splendor*, 1612. In the New York Public Library.

By courtesy of the New York Public Library, Rare Book Division, Astor, Lenox and Tilden Foundations

Although the garden is designed around a central axis, the stream is not used centrally but is led about the garden in order to take maximum advantage of its force. Unlike the less copious stream of the Villa Lante garden, which quietly emphasizes the central axis, the Tivoli stream is ostentatious. The Villa d'Este is, in fact, a spectacular permanent theatrical performance meant to astonish and impress the multitude. A different impression is given by the Boboli Gardens of the Pitti Palace at Florence (1550). Though, like the Villa d'Este gardens, they were designed for a crowd—specifically, for state functions—they are not dramatic in themselves. Unless used ceremonially, they are lifeless and arid. The ruined garden associated with, though detached from, the Orsini Castle at Bomarzo is a remarkable aberration probably influenced by accounts of visits to the Far East by a locally born traveller, Biagio Sinibaldi. Its original layout consisted of a grove in which were concealed the stone giants and strange monsters that now astonish visitors.

Flowers were extensively used in most Italian gardens, but because of the shortness of their season they could not be the principal feature. Beds were divided into decorative geometric compartments by trimmed herbs, rosemary, lavender, or box. In general, more emphasis was given to evergreens; ilex, cypress, laurel, and ivy gave shade and were an enduring contrast to stonework.

G. Tomish—Photo Researchers

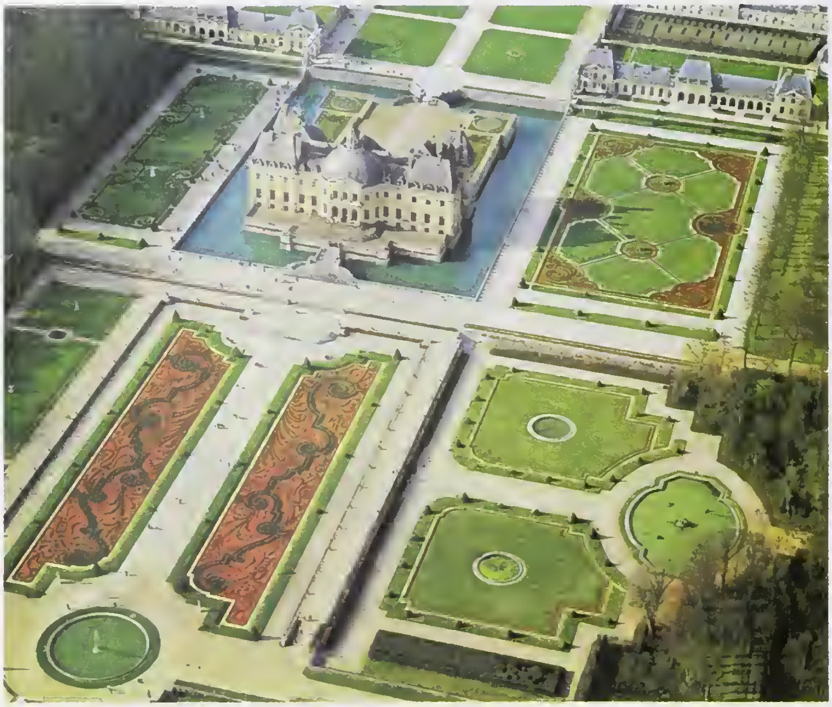
Elaborate hillside fountain in the gardens of the Villa d'Este at Tivoli, Italy, mid-16th century.

Influence of the Italian garden in France

17th- and 18th-century French. The French invasions of Italy in the last quarter of the 16th and first quarter of the 17th centuries introduced to France the idioms of the Italian garden. The first garden coordinated with dwelling appeared at the Château of Anet (1547–56) and was designed by the architect Philibert Delorme; but, despite its evident sophistication, it remained an inward-looking, essentially medieval garden. The first sign of prolongation and calculated extension of vision beyond the garden proper appeared in the grounds of Dampierre. Here the moat that formerly surrounded French castles became an ornamental body of water on one side and a decorative canal on the other. Both aspects of the new garden design—coordination with the dwelling and extension along a central axis—were united at the château of Richelieu (1631) and later at Vaux-le-Vicomte (completed 1661), the château of Nicolas Fouquet, the minister of finance. On Fouquet's fall in the mid-17th century, his team of artists—which included the landscape designer André Le Nôtre—was taken over by the young Louis XIV, and the gardens of Versailles were begun.

Typical French classical-baroque gardens

The French version of the Italian garden was created in the plain of north France, which largely conditioned the manner of its development. The array of steep terraces linked by stairways, which characterized the Villa d'Este and many others, was predominant in France only at St. Germain-en-Laye, where the steep site permitted it. Elsewhere, grandeur on the scale that competitive pride demanded was achieved by extraordinary extension: an axial development suggesting a domain coextensive with the world. The French 17th-century garden, a manifestation of Baroque taste, required variety as well as unlimited vista and achieved it with fountains, parterres, and lesser gardens disposed within the boscages (wooded enclosures) that flanked the central axis. These hidden gardens were the successors of the *giardini segreti* of the Italians but had a different function; they were not retreats for private contemplation or intimate conversation but the setting for ingenious theatrical entr'actes. Distinctively French was the unified and elaborate treatment of the compartmentalized garden beds, which the Italians had made in a variety of forms. These *compartiments de broderie* were arabesques, sometimes of box edging and flowers but more often of coloured stones and sand. The Persians had copied their flower gardens on carpets and taken them indoors; the French laid out their grounds in the form of carpets. The French garden was marked by a ruthlessly



Parterre de broderie composed of box, brick dust, earth, and sand, at the château of Vaux-le-Vicomte, France, a late 19th-century re-creation by Henri Duchêne based upon late 17th-century designs by André Le Nôtre.
Alain Perceval, Paris

logical extension of practices that had been empirically evolved in Italy.

French cultural dominance of Europe in the early 18th century led to an almost universal adoption of Versailles as the model for palatial gardens. Even at Naples, where the gardens of Poggio Reale had astonished the invading French in the late 15th century, a vast layout inspired by the axial extent of Versailles was developed at Caserta; and, as far away as Peter the Great's Peterhof in Russia, a pseudo-Versailles was laid out by the French gardener Alexandre-Jean-Baptiste Le Blond. Impressive exercises in the same manner were carried out in Germany and Austria. In Holland, also, the example of the French garden was irresistible, although local conditions and national temperament led to regional variation. Because Dutch canals were busy highways, they generally flanked gardens rather than constituted the main axis. No luxury in Holland, water was less extravagantly used than in drier, hotter climates. Moreover, fountains were less common because the absence of high ground required that they be power driven. Because stone was scarce, terraces were usually held by turf banks rather than by retaining walls, and sculpture was often of lead. Another sculpture typical of the Dutch garden was topiary: trees and shrubs were trained, cut, and trimmed into sculptural, ornamental shapes. Social conditions made the extension of a geometric garden easy, for a man-made landscape already existed in the intensively cultivated Netherlands. In Spain, aridity as well as Islāmic tradition perpetuated the patio garden, a room of air and shade in the Greek peristyle tradition. Although a famous layout in the French style was made on high ground at La Granja, where the cooler air and ample water made it acceptable, the classical extension garden remained basically alien to the Iberian Peninsula.

By courtesy of the City of Birmingham Museum and Art Gallery, Birmingham, England, photograph, Reilly and Constantine

"Garden at Utrecht," by Issac de Moucheron (1667–1744), showing the Dutch preference for high, trimmed hedges and alleyways. In the City of Birmingham Museum and Art Gallery, Birmingham, England.

17th- and 18th-century English. The Italian pronouncement "things planted should reflect the shape of things built" had ensured that gardens were essentially open-air buildings and the making of them the province of architects. Before the 18th century, geometric regularity had been applied in great details of design and in small. England was committed to a version of the French geometric extension garden but with an emphasis on English grass lawns and gravel walks. Whereas the typical French vista was along the main axis, with subordinate vistas at right angles to it, in the two most influential gardens in England, St. James's and Hampton Court, the vistas sprang like the rays of the sun from a semi-circle. With the accession of William and Mary (1689–1702), Dutch influence led to widespread use of topiaried yew and box.

In 18th-century England, people became increasingly aware of the natural world. Rather than imposing their man-made geometric order on the natural world, they began to adjust to it. Literary men, notably Alexander Pope and Joseph Addison, began to question the propriety of trees being carved into artificial shapes as substitutes for masonry and to advocate the restoration of free forms.

Revolt against the "artificial" symmetrical garden

The man who led the revolt against the "artificial," symmetrical garden style was the painter and architect William Kent, the factotum of the Earl of Burlington. Together Burlington and Kent created at Chiswick House (1734) a garden with a meandering stream and an "irregular" path. As the writer Horatio Walpole put it, Kent's "principle was that nature abhors a straight line." The process of relaxing the garden's architectural discipline advanced with speed. At Stowe, Buckinghamshire, the original enclosed, geometrical garden was amended over the years until a totally different, "irregular" formality was achieved. Trees, for example, were allowed to assume their natural forms, and a large expanse of water was redesigned into two irregularly shaped lakes.

The use of the ha-ha, or sunken fence, to create and at the same time conceal the physical division between garden and contiguous park grounds (a division needed to keep grazing animals out of the garden) was a major step in the creation of the new, "natural" garden. Walpole explains the purpose of the visual unification:

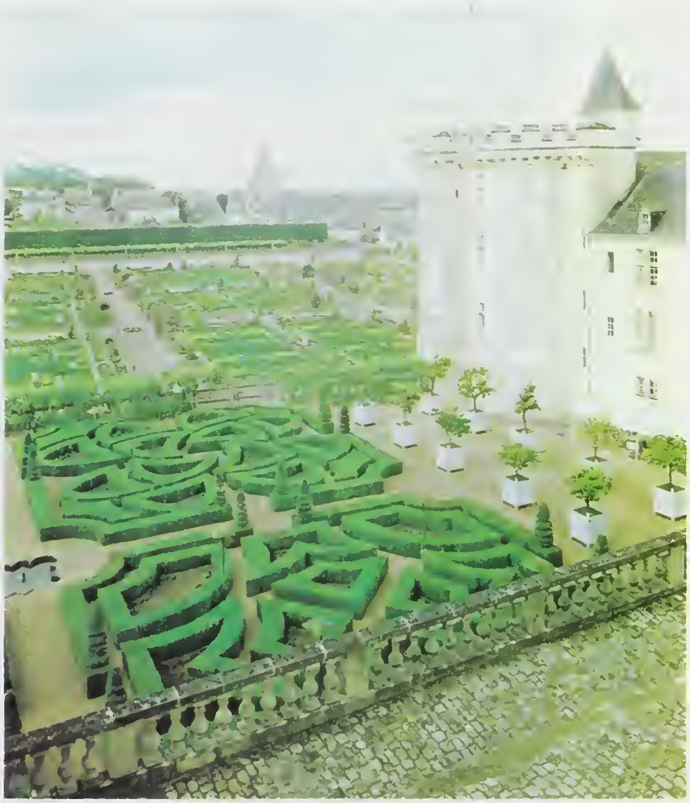
> The contiguous ground of the park without the sunk fence was to be harmonized with the lawn within; and the garden in its turn was to be set free from its prime regularity, that it might assort with the wilder country without.

The "natural" English gardens of the 18th century

The face of the "country without" was altered by the rage that afflicted the English nobility for planting vast areas of trees. Much of England was covered with new parks, traversed by rides and avenues that primarily were conceived as visual extensions of the garden paths. The unification of park and garden was virtually completed by Lancelot "Capability" Brown (1715–83) by the simple expedient of making the garden into a park. "Capability" (so-called because he always spoke of a place as having "capabilities of improvement") developed the current aesthetic that an undulating line was "natural" and that it was the "line of beauty" by using little statuary and few buildings and concentrating on designing landscapes according to nature's harmonies and gradients. His landscapes consist of expanses of grass, irregularly shaped bodies of water, and trees placed singly and in clumps.

Although the adherents of the new English school of garden design were in agreement in their abhorrence of the straight, classical line and the geometrically ordered garden, they did not agree on what the natural garden should be. Unlike Brown, for example, the taste for the romantic and the literary led many to seek inspiration in the dramatic and the bizarre, in the remote past, and in remote, exotic places. The Brownian style was strongly challenged, for example, by the "Picturesque" school, led by Sir Uvedale Price and the artist-parson William Gilpin, who argued, quite correctly, that the "naturalism" of the Brownians was no less unnatural than the geometric regularity of Le Nôtre's Versailles and that sudden declivities, rocky chasms, and rotting tree trunks (all deliberately designed) were more proper for the natural garden than were enormous, undulating meadows accented with tight clumps of thickly planted trees. Another school of opinion created what might be called the English garden of poetic bric-a-brac. The aim in this garden was to create an air of accident and surprise and to arouse varied sensations (solemnity, sublimity, terror) in the viewer—sensations evoked by associations with the remote in time and space. Wandering through the grounds one came upon classical statues, urns, and temples; Gothic ruins, ivy-covered and inhabited by owls; or Chinese pagodas and bridges. After Horatio Walpole recorded the first appearance of chinoiserie at Wroxton in 1753 (a garden no doubt laid out some years before), "Chinese" and Gothic details were featured, together with classic temples, in most fashionable grounds.

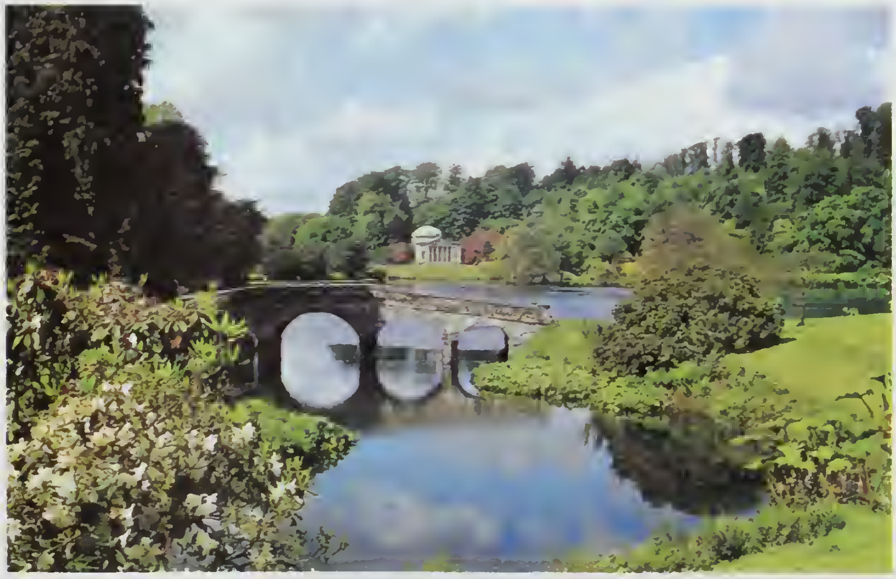
By 1760 the enthusiasm for this style had diminished in England; but in Europe the poetic bric-a-brac garden (*le jardin anglo-chinois*, or *le jardin anglais*, as the French



Reconstruction of the 16th-century gardens at Villandry, in the Loire Valley, France.

Edwin Smith

Noel Habgood Eastbourne



Idealized natural landscape, gardens at Stourhead, Wiltshire, designed by the owner, Henry Hoare, 18th century.

English influence on European gardens

called it) was almost as widely emulated as Versailles had been. In Italy, for example, Renaissance gardens were destroyed to make way for the new fashion, as at the Villa Mansi near Lucca. In France, the sculpted group of "Apollo Attended by the Nymphs" was removed from the classical "Grotto of Thetis" on the terrace of Versailles to a secluded boscage garden, where it was housed under ornamental "Turkish" tents; eventually it was moved from there to a simulated rocky cavern in the *jardin anglais* of the Petit Trianon. The *jardin anglais* was to be found even at Queluz in Portugal and in the Potsdam garden of Frederick the Great of Prussia.

19th century. Increasing world trade and travel brought to late 18th-century Europe a flood of exotic plants whose period of flowering greatly extended the potential season of the flower garden. Although the emphasis in Italian Renaissance gardens, in the classical Baroque gardens of France, in the lawns and gravelled walks of 17th-century England, and in the Brownian park garden was upon design, they had rarely been totally without flowers. In most gardens flowers were grown, sometimes in great numbers and variety; but flower gardens in the modern sense were limited to cottages, to small town gardens, and to relatively small enclosures within larger gardens. The accessibility of new plants, together with avidity for new experience and a high-minded concern with natural science, not only gave renewed life to the flower garden but was the first step toward the evolution of the garden from work of art to museum of plants. A compromise between the new flower garden and the Brownian park was effected by Humphry Repton. He was largely responsible for popularizing the open terrace overlooking the park, which frankly admitted the different functions of park and garden and also emphasized their stylistic disharmony. The plant collectors' garden, or "gardenesque" style, was most strongly advanced by J.C. Loudon in the mid-19th century. Loudon urged that garden making be taken out of the hands of the architect, the painter, and the cultivated dilettante and left to the professional plantsman.

The undiscerning use of the new palette that importation and plant breeding had made available was so patently an aesthetic disaster that by the end of the 19th century attempts were made to break its hold. The architect Sir Reginald Blomfield advocated a return to the formal garden, but to this, insofar as it required dressed stonework, there were economic objections. More successful and more in tune with the escapist needs of the increasing number of urban dwellers were the teaching and practice of William Robinson, who attacked both the old ceremonial garden and the collectors' garden with equal vigour and preached that botany was a science, but gardening was an art. Under his leadership a more critical awareness was brought to the planning and planting of gardens. His own garden at Gravetye Manor demonstrated that plants look best where they grow best and that they should be allowed to develop their natural forms. Adapting Robinson's principles, Gertrude Jekyll applied the cult of free forms over a substructure of concealed architectural regularity, bringing the art of the flower garden to its highest point.

Early gardens in the United States

In North America, where for a long time most men were preoccupied with making a world, not a garden, ornamental gardens were slow to take hold. In the gardens that did exist, the rectilinear style popular in late 17th- and early 18th-century Europe persisted well into the 18th century—perhaps because it met man's psychological need to feel he could master a world that was still largely untamed. The town gardens of Williamsburg (begun in 1698) were typical of the Anglo-Dutch urban gardens that were being attacked everywhere in 18th-century Europe except Holland. And Belmont, in Pennsylvania, was laid out as late as the 1870s with mazes, topiary, and statues, in a style that would have been popular in England about two centuries before.

Although garden improvers set up in business in the United States, there is no evidence that they prospered until the 19th century, when one hears of André Parmentier, a Belgian, who worked on Hosack's estate at Hyde Park and then of A.J. Downing, a successful protagonist of the gardenesque, who was succeeded by Calvert Vaux and Frederick Law Olmsted (the latter the originator of the title and profession of landscape architect), the planners of Central Park (begun 1857) in New York City and of public parks throughout the country.

The eclecticism of the 19th century was universal in the Western world. Besides the gardens that were fundamentally Reptonian—that is, an attempted compromise between the Brownian park garden and the Loudonian flower garden—gardens of almost every conceivable style were copied; designing teams such as Sir Charles Barry, the architect, and William Eden Nesfield, the painter, in England, for example, produced Italianate parterres as well as winding paths through thickets.

Modern. A sense of history still plays a part in 20th-century gardening. The desire to maintain and reproduce old gardens, such as the reconstruction of the 16th-century gardens of Villandry in France and the colonial gardens of Williamsburg in the United States, is not peculiarly modern (similar things were done in the 19th century); but, as man increasingly needs the reassurance of the past, it is likely to be, within limits, an accelerating process.

By courtesy of Dr Edward S Ayensu, Smithsonian Institution

Parque del Este, Caracas, designed by Roberto Burle Marx, 1959.

Attempts to create a distinctive modern idiom are rare. Gardens large by modern standards are still made, in styles that vary from a version of the grand early 18th-century manner at Anglesey Abbey in Cambridgeshire to an inflated Jekyllism crossed with gardenesque at Bodnant near Conway. An air either of controlled wilderness or of slightly run-to-seed orderliness is preferred. Modern public gardens, which have evolved from the large private gardens of the past, seek instant popular applause for the quantity and brightness of their flowers. In Brazil Roberto Burle Marx uses tropical materials to give an air of contemporaneity to traditional modes of design. Gardens frequently reflect Japanese influence, particularly in America.

Functional planning

Most characteristic of the 20th century is functional planning, in which landscape architects concentrate upon the arrangement of open spaces surrounding factories, offices, communal dwellings, and arterial roads. The aim of such planning has been to provide, at best, a satisfactory setting for the practical aspects of living. It is gardening only in the negative, "tidying up" sense, with little concern for the traditional garden purpose of awakening delight. So starved has the spirit of those living in heavily populated regions become, however, that demands grow more and more insistent for gardening in the positive sense—for environmental planning with a chief goal not of facilitating man's economic activities but of refreshing his spirit.

NON-WESTERN

Chinese. Western gardens for many centuries were architectural, functioning as open-air rooms and demonstrating Western man's insistence on physical control of his environment. Because of a different philosophical approach, Oriental gardens are of a totally different type.

China—which is to Eastern civilization what Egypt, Greece, and Rome are to Western—practiced at the beginning of its history an animist form of religion. The sky, mountains, seas, rivers, and rocks were thought to be the materialization of spirits whom men regarded as their fellow inhabitants in a crowded world. Such a belief emphasized the importance of good manners toward the world of nature as well as toward the world of men. Against this background the Chinese philosopher Lao-tzu taught the quietist philosophy of Taoism, which held that one should integrate oneself with the rhythms of life; Confucius preached moderation as a means of attaining spiritual calm; and the teaching of Buddha elevated the attainment of calm to a mystical plane.

Such a history of thought led the Chinese to take keen pleasure in the calm landscape of the remote countryside. Because of the physical difficulty of frequent visits to the sources of such delight, the Chinese recorded them in landscape paintings and made three-dimensional imitations of them near at hand. Their gardens were therefore representational, sometimes direct but more often by substitution, making use of similar means to recreate the emotions that choice natural landscapes evoked. The kind of landscape that appealed was generally of a balanced sort; for the Chinese had discovered the principle of complementary forms, of male and female, of upright and recumbent, rough and smooth, mountain and plain, rocks and water, from which the classic harmonies were created. The principle of scroll painting, whereby the landscape is exposed not in one but in a continual succession of views, was applied also in gardens; and grounds were arranged so that one passed pleasantly from viewpoint to viewpoint, each calculated to give a different pleasure appropriate to its situation. A refined and expectant aestheticism, which their philosophy had inculcated, taught the Chinese to ignore nothing that would prepare the mind for the reception of such experiences, and every turn of path and slope of ground was carefully calculated to induce the suitable attitude. As the garden was in effect a complex of linked, related, but distinct sensations, seats and shelters were situated at chosen spots so that the pleasures that had been meticulously prepared for could be quietly savoured. Kiosks and pavilions were built at places where the dawn could best be watched or where the moonlight shone on the water or where autumn foliage was seen to advantage or where the wind made music in the bamboos. Such gardens were intended not for displays of wealth and magnificence to impress the multitude but for the delectation of the owner, who felt his own character enhanced by his capacity for refined sensation and sensitive perception and who chose friends to share these pleasures with the same discernment as he had exercised in planning his garden.

The typical Chinese garden

Based on natural scenery, Chinese gardens avoided symmetry. Rather than dominating the landscape, the many buildings in the garden "grew up" as the land dictated. A fanciful variety of design, curving roof lines, and absence of walls on one or on all sides brought these structures into harmony with the trees around them. Sometimes they were given the rustic representational character of a fisherman's hut or hermit's retreat. Bridges were often copied from the most primitive rough timber or stone-slab raised pathways. Rocks gathered from great distances became a universal decorative feature, and a high connois-

By courtesy of the St. Louis Art Museum W.K. Bixby Oriental Art Fund

Harmony and intimacy of the Chinese garden. "Enjoyment of the Chrysanthemum Flowers," ink and colours on paper by Hua Yen, 1753. In the St. Louis Art Museum.

seurship developed in connection with their colour, shape, and placement.

Although the troubled 20th century has largely destroyed the old gardens, paintings and detailed descriptions of them dating from the Sung dynasty (AD 960–1279) reveal a remarkable historical consistency. Nearly all the characteristic features of the classic Chinese garden—man-made hills, carefully chosen and placed rocks, meanders and cascades of water, the island and the bridge—were present from the earliest times.

Influence of Chinese gardens on the West

Chinese gardens were made known to the West by Marco Polo, who described the palace grounds of the last Sung emperors, during whose reign the arts were at their most refined. Other accounts reached Europe from time to time but had little immediate effect except at Bomarzo, the Mannerist Italian garden that had no successors. In the 17th century, the English diplomat and essayist Sir William Temple, sufficiently familiar with travellers' tales to describe the Chinese principle of irregularity and hidden symmetry, helped prepare the English mind for the revolution in garden design of the second quarter of the 18th century. Chinese example was not the sole or the most important source of the new English garden; but the account of Father Attiret, a Jesuit at the Manchu court, published in France in 1747 and in England five years later, promoted the use of Chinese ornament in such gardens as Kew and Wroxton and hastened the "irregularizing" of grounds. The famous *Dissertation on Oriental Gardening* by the English architect Sir William Chambers (1772) was a fanciful account intended to further the current revolt in England against the almost universal Brownian park garden. Influence of the West on Chinese gardens was slight. Elaborate fountain works, Baroque garden pavilions, and mazes—all of which the Jesuits made for the imperial garden at Yüan Ming Yüan—took no root in Chinese culture. Not until the 20th century did European regularity occasionally become evident near the Chinese dwelling; at the same time, improved Western hybrids of plant species that had originated in the East appeared in China.

Japanese. Chinese culture permeated the Far East and, by way of Korea, infiltrated Japan. By the year AD 1000 Japan was already developing a distinctive national art best described as a stylized, ritualistic version of the Chinese. The typical early Japanese garden lay to the south of the dwelling and consisted of a narrow pond or lake orientated through its longer axis and containing an island. At the north end of the pond was an artificial hill from which a secondary stream descended in a cascade. These stereotyped gardens of the Heian period (AD 794–1185) show by their careful reproduction of magical detail that they derive from a single prototype—certainly Chinese. Variation entered only through the individual particularities of the site and the detailed handling of stones and trees.

William G. Froelich, Jr

Pond and moss-covered bridge, Katsura Imperial Gardens, Kyōto, Japan.

Creativity began to replace imitation in the Kamakura period (1192–1333). Although there were many subsidiary styles, gardens were broadly classified, according to terrain, as either hill or flat. The hill garden, consisting of hills and ponds, came to be associated with Mt. Fuji, the mountain of ideal form. The flat garden represented a surface of water—lake or sea—with its adjacent shores and islands. Since the scale was so small—a heap of earth 30 feet (nine metres) high representing a mountain and a half-acre (0.2-hectare) pond, an arm of the sea—the intention was to reproduce the spirit rather than the features of the chosen landscape. Association and symbolism thus played a major role in the creation and appreciation of these gardens.

Types of Japanese gardens

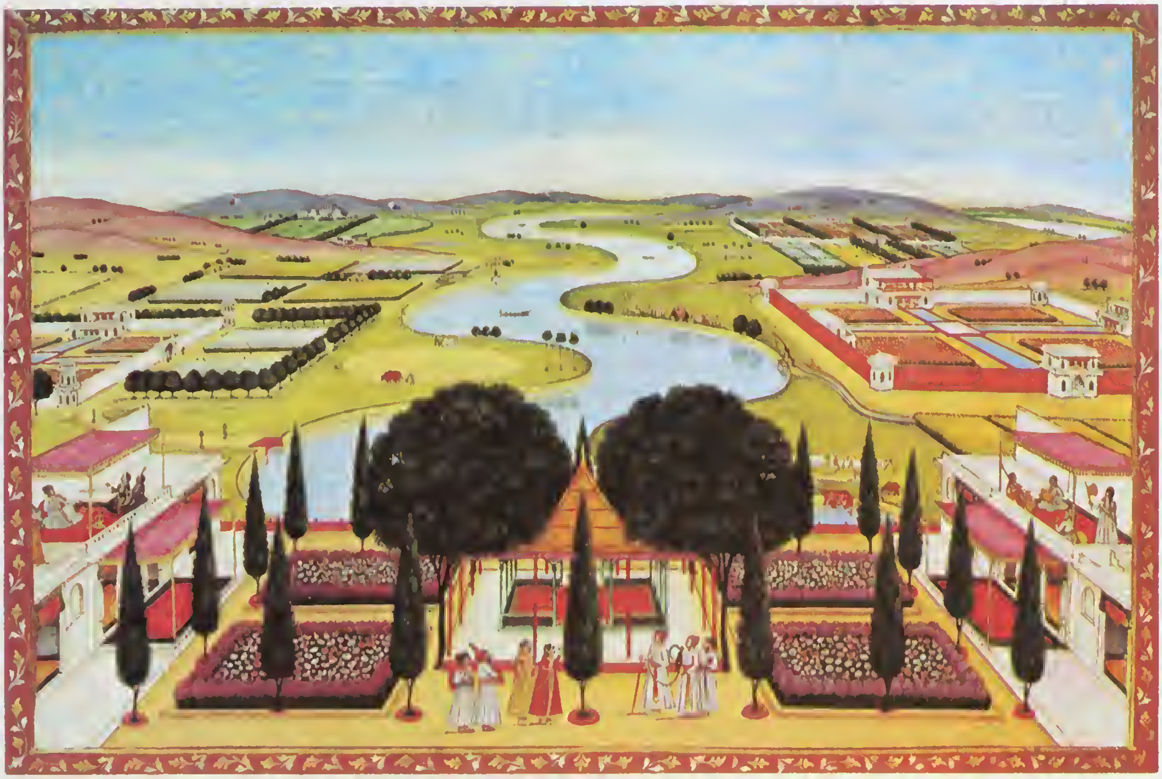
The scaling down of landscapes to garden size was logically continued to the point where miniature gardens were made in trays as small as a foot square containing lakes, streams, islands, hills, bridges, garden houses, and real trees painstakingly cultivated to an appropriate scale. These small, portable gardens reflected the extreme of the picturesque tradition of Oriental gardening.

Abstract and tea gardens

Two characteristic Japanese styles are the abstract garden and the tea garden. The most famous example of the former is the garden of the Ryōan-ji in Kyōto, where an area about the size of a tennis court is covered with raked sand and set with 15 stones divided into five groups. If anything is represented here, it is some rocky islets in a sea, but the appeal of the garden lies essentially in the charm of its relationships. The Japanese tea garden grew out of an esoteric ritual originated in China and connected with the taking of tea. The tea cult, which flourished from the 14th to the end of the 16th centuries, was calculated to instill humility, restraint, sensibility, and other cognate virtues. The gardens through which the guests approached the teahouse were governed by severe rules of design intended to create an appropriate spiritual atmosphere, such as the "lonely precincts of a secluded mountain shrine" or "a landscape in clouded moonlight, with a half-gloom between the trees" or any mood "in harmony with the spirit of tea." Even the precise number and arrangement of nails in the teahouse door were specified.

The use of stones

The Japanese fondness for systematization led them to classify garden treatment as well as subject. Three standard treatments were recognized: the elaborate, the moderate, and the modest. Once the degree of finish was determined, certain rules were followed to preserve consistency. The Taoist doctrine of complementary forms was at the root of much Japanese design, but the cult of stones is also central to Japanese gardening. The nine stones, five standing and four recumbent, used in Buddhist gardens were symbols of the nine spirits of the Buddhist pantheon; the shapes and postures chosen were presumed to have a relationship with the character and history of the persons represented. Sacred associations played a part in profane gardens as well. It was regarded as inauspicious, for example, if three stones, the "Guardian Stone," the "Stone of Adoration," and the "Stone of the Two Deities" (or the "Stone of Completeness"), were not present. In addition to the sacred symbols, a whole armoury of poetic associations and symbols grew up, and stones, according to their shape and use, acquired such names as "Torrent-breaking Stone," "Recumbent Ox Stone," "Propitious Cloud Stone," and "Seagull-resting Stone." Beyond what they represented, stones were part of an aesthetic design and had to be placed so that their positions appeared natural and their relationships harmonious. The concentration of the interest on such detail as the shape of a rock or the moss on a stone lantern led at times to an overemphatic picturesqueness and an accumulation of minor features that, to Occidental eyes accustomed to a more general survey,



Architecturally planned park landscape for an imperial residence, containing an artificially meandering river and pavilions. An example of the ideal landscapes of Mughal India. Miniature from a 17th- or 18th-century album. In the Staatliche Museen zu Berlin.

By courtesy of the Staatliche Museen zu Berlin

may seem cluttered and restless. Nevertheless, Japanese gardening has had and continues to have an influence on the gardens of the West, particularly in the United States. The influence appears not so much in direct imitation of Japanese themes as in the selection and presentation of detail.

Indian. The influence of Chinese culture throughout the East was such that other indigenous cultures usually succumbed to it; but India was an exception. Western garden styles were introduced into northern India first through contact with Iranian culture, then by the invasion of Alexander the Great and the subsequent Hellenistic influence, and, finally, by the invading Mughals, who introduced the Islāmic garden.

In southern India and in Ceylon elaborate gardens existed before the birth of the Buddha (563? BC). Beneath a tree in such a garden—containing baths, lotus-covered pools, trees, and beds of flowers—the Buddha himself was said to have been born. Anciently worshipped by the Hindus, trees thus acquired an additional sanctity. Buddhist temples were associated with gardens whose purpose was to promote contemplation and whose preferred sites were therefore away from cities.

African, Oceanic, and pre-Columbian. The African cultures beyond European and Asiatic influence did not evolve pleasure gardens, although in their more settled societies a beginning had perhaps been made. Nor is more than a love of flowers and a casual cultivation of decorative plants recorded of the Oceanic peoples; but of the Aztecs of Mexico and the Incas of Peru the conquistadores reported elaborate gardens with terraced hills, groves, fountains, and ornamental ponds that were essentially royal pleasure grounds, reflecting a need for private solace and public display not unlike contemporary gardens in the West. (D.P.Cl.)

BIBLIOGRAPHY. The 19th and 20th centuries have seen a spate of gardening literature, both of the how-to-do-it variety and of specialized studies of national types. General works include DEREK CLIFFORD, *A History of Garden Design*, 2nd ed. (1966), an attempt to relate garden design to cultural history; and MARIE LUISE GOTHEIN, *A History of Garden Art*, 2 vol. (1928; originally published in German, 1914), the most valuable detailed general history.

National studies: (*America*): WILLIAM COBBETT, *The American Gardener* (1819); A.J. DOWNING, *A Treatise on the Theory and Practice of Landscape Gardening* (1841, reissued 1977); FREDERICK L. OLMSTED, *Forty Years of Landscape Architecture*, 2 vol. (1922–28; reprinted in 1 vol., 1970). (*Brazil*): P.M. BARDI, *The Tropical Gardens of Burle Marx* (1964). (*China*): OSVALD SIREN, *Gardens of China* (1949). (*Great Britain*): ALICIA AMHERST, *A History of Gardening in England*, 3rd ed. (1910); SIR REGINALD BLOMFIELD and F. INIGO THOMAS, *The Formal Garden in England*, 3rd ed. (1901, reprinted 1972); MILES HADFIELD, *Gardening in Britain* (1960); CHRISTOPHER HUSSEY, *English Gardens and Landscapes, 1700–1750* (1967); EDWARD S. HYAMS, *The English Garden* (1964), a well-informed guidebook. (*India*): C.M. VILLIERS-STUART, *Gardens of the Great Mughals* (1913, reprinted 1979); SYLVIA CROWE, *The Gardens of Mughal India: A History and a Guide* (1972). (*Italy*): J.C. SHEPHERD and G.A. JELLICOE, *Italian Gardens of the Renaissance*, 3rd ed. (1966), the best book; GFORGINA MASSON, *Italian Gardens*, new ed. (1966), a well-informed but not always accurate guidebook that lacks historical coherence; H. INIGO TRIGG, *The Art of Garden Design in Italy* (1906), the source of many 20th-century reproductions of Italian gardens; EDITH WHARTON, *Italian Villas and Their Gardens* (1904, reissued 1976), still worth attention. (*Japan*): LORAINE E. KUCK, *The World of the Japanese Garden* (1968, reissued 1980), the best book; JOSIAH CONDOR, *Landscape Gardening in Japan* (1893); JIRŌ HARADA, *The Gardens of Japan* (1928); TAKEJI IWAMIYA, *Imperial Gardens of Japan* (1970). (*Persia*): DONALD N. WILBER, *Persian Gardens and Garden Pavilions*, 2nd ed. (1979). (*Spain*): C.M. VILLIERS-STUART, *Spanish Gardens: Their History, Types and Features* (1929).

Modern works: PETER F. SHEPHEARD, *Modern Gardens* (1953); THOMAS D. CHURCH, *Gardens Are for People*, 2nd ed. (1983); JAMES C. ROSE, *Creative Gardens* (1958); GARRETT ECKBO, *The Art of Home Landscaping* (1956), and *The Landscape We See* (1969); PETER COATS, *Great Gardens of the Western World* (1963); ELIZABETH B. KASSLER, *Modern Gardens and the Landscape*, rev. ed. (1984); SUSAN and GEOFFREY JELLICOE, *Modern Private Gardens* (1968); GEOFFREY JELLICOE, *Studies in Landscape Design* (1960); LAWRENCE HALPRIN, *Cities* (1963, reissued 1972); EDMUND N. BACON, *Design of Cities*, rev. ed. (1974); IAN L. MCHARG, *Design with Nature* (1969, reissued 1971).

(D.P.Cl./G.Ec.)

Gardening and Horticulture

Gardening belongs both to art and to science. It is an artistic activity in that it deals with the grouping of plants in harmonious or pleasing arrangements. It is scientific in its concern with the techniques of cultivating plants and producing satisfactory growth.

Horticulture is the more comprehensive term, embracing many forms of production from the soil. In common usage, and in the sense applied here, it refers to commercial gardening. It thus includes the growing of flowers, fruits, and vegetables as crops for profit, as in market gardening; the breeding of plants for sale, as in nursery gardening; and landscape architecture and design. The seed industry is also closely associated with horticulture.

Because plants are often grown in conditions markedly different from those of their natural environment, it is necessary to apply to their cultivation techniques derived from plant physiology, chemistry, and botany, modified by the experience of the planter. The basic principles involved in growing plants are the same in all parts of the world, but the practice naturally needs much adaptation to local conditions. The three main topics covered in this article, gardening, houseplants, and horticulture, thus have much in common, being closely linked in historical development and principles. The differences between them arise from size and scale.

For the main history of garden development, see the article GARDEN AND LANDSCAPE DESIGN.

This article is divided into the following sections:

Gardening

THE NATURE OF GARDENING

Gardening in its ornamental sense needs a certain level of civilization before it can flourish. Wherever that level has been attained, in all parts of the world and at all periods, people have made efforts to shape their environment into an attractive display. The instinct and even enthusiasm for gardening thus appear to arise from some primitive response to nature, engendering a wish to produce growth and harmony in a creative partnership with it.

It is possible to be merely an admiring spectator of gardens. However, most people who cultivate a domestic plot also derive satisfaction from involvement in the processes of tending plants. They find that the necessary attention to the seasonal changes, and to the myriad small "events" in any shrubbery or herbaceous border, improves their understanding and appreciation of gardens in general.

The upsurge of interest in gardening

A phenomenal upsurge of interest in gardening began in Western countries after World War II. A lawn with flower beds and perhaps a vegetable patch has become a sought-after advantage to home ownership. The increased interest produced an unprecedented expansion of business among horticultural suppliers, nurseries, garden centres, and seedsmen. Books, journals, and newspaper columns on garden practice have found an eager readership, while television and radio programs on the subject have achieved a dedicated following.

Several reasons for this expansion suggest themselves. Increased leisure in the industrial nations gives more people the opportunity to enjoy this relaxing pursuit. The increased public appetite for self-sufficiency in basic skills also encourages people to take up the spade. In the kitchen, the homegrown potato or ear of sweet corn rewards the gardener with a sense of achievement, as well as with flavour superior to that of store-bought produce. An increased awareness of threats to the natural environment and the drabness of many inner cities stir some people to cultivate the greenery and colour around their own doorsteps. The bustle of 20th-century life leads more individuals to rediscover the age-old tranquillity of gardens.

The varied appeal of gardening. The attractions of gardening are many and various and, to a degree perhaps unique among the arts and crafts, may be experienced by any age group and at all levels of ambition. At its most elemental, but not least valuable, the gardening experience begins with the child's wonder that a packet of seeds will produce a charming festival of colour. At the adult level it can be as simple as helping to raise a good and edible carrot, and it can give rise to almost parental pride. At higher levels of appreciation, it involves an understanding of the complexity of the gardening process, equivalent to a chess game with nature, because the variables are so many.

The gardening experience may involve visiting some of the world's great gardens at different seasons to see the relation of individual groups of plants, trees, and shrubs to the whole design; to study the positioning of plants in terms of their colour, texture, and weight of leaf or blossom; and to appreciate the use of special features such as ponds or watercourses, pavilions, or rockeries. Garden visiting on an international scale provides an opportunity to understand the broad cultural influences, as well as the variations in climate and soil, that have resulted in so many different approaches to garden making.

The appeal of gardening is thus multifaceted and wide in range. The garden is often the only place where someone without special training can exercise creative impulses as designer, artist, technician, and scientific observer. In addition, many find it a relaxing and therapeutic pursuit. It is not surprising that the garden, accorded respect as a part of nature and a place of contemplation, holds a special place in the spiritual life of many.

Practical and spiritual aspects of gardening are shown in an impressive body of literature. In Western countries manuals of instruction date to classical Greece and Rome. Images of plants and gardens are profuse in the works of the major poets, from Virgil to Shakespeare, and on to some of the moderns.

Another of gardening's attractions is that up to a certain level it is a simple craft to learn. The beginner can produce pleasing results without the exacting studies and practice required by, for example, painting or music. Gardens are also forgiving to the inexperienced to a certain degree. Nature's exuberance will cover up minor errors or short periods of neglect, so gardening is an art practiced in a relatively nonjudgmental atmosphere. While tolerant in many respects, nature does, however, present firm reminders that all gardening takes place within a framework of natural law; and one important aspect of the study of the craft is to learn which of these primal rules are imperatives and which may be stretched.

Control and cooperation. Large areas of gardening development and mastery have concentrated on persuading plants to achieve what they would not have done if left in the wild and therefore "natural" state. Gardens at all times have been created through a good deal of control and what might be called interference. The gardener attends to a number of basic processes: combating weeds and pests; using space to allay the competition between plants; attending to feeding, watering, and pruning; and conditioning the soil. Above this fundamental level, the gardener assesses and accommodates the unique complex of temperature, wind, rainfall, sunlight, and shade found within his own garden boundaries. A major part of the fascination of gardening is that in problems and potential no one garden is quite like another; and it is in finding the most imaginative solutions to challenges that the gardener demonstrates artistry and finds the subtler levels of satisfaction.

Basic processes of control

Different aesthetics require different balances between controlling nature and cooperating with its requirements. The degree of control depends on the gardener's objective, the theme and identity he is aiming to create. For example, the English wild woodland style of gardening in the mid-19th century dispensed with controls after planting, and any interference, such as pruning, would have been misplaced. At the other extreme is the Japanese dry-landscape garden, beautifully composed of rock and raked pebbles. The artistic control in this type of garden is so firm and refined that the intrusion of a single "natural" weed would spoil the effect.

Choice of plants. The need for cooperation with nature is probably most felt by the amateur gardener in choosing the plants he wants to grow. The range of plants available to the modern gardener is remarkably rich, and new varieties are constantly being offered by nurseries. Most of the shrubs and flowers used in the Western world are descendants of plants imported from other countries. Because they are nonnative, they present the gardener with some of his most interesting problems, but also with the possibility of an enhanced display. Plants that originated in subtropical regions, for example, are naturally more sensitive to frost. Some, like rhododendrons or azaleas, originated in an acid soil, mainly composed of leaf mold. Consequently they will not thrive in a chalky or alkaline soil. Plant breeding continues to improve the adaptability of such exotic plants, but the more closely the new habitat resembles the original, the better the plant will flourish. Manuals offer solutions to most such problems, and the true gardener will always enjoy finding his own. In such experiments, he may best experience his work as part of the historical tradition of gardening.

HISTORICAL BACKGROUND

Early history. Western gardening had its origins in Egypt some 4,000 years ago. As the style spread it was changed and adapted to different localities and climates, but its essentials remained those of disciplined lines and groupings of plants, usually in walled enclosures. Gardening was introduced into Europe through the expansion of Roman rule and, second, by way of the spread of Islām into Spain. Though clear evidence is lacking, it is presumed that Roman villas outside the confines of Italy contained native and imported plants, hedges, fruit trees, and vines, in addition to herbs for medicinal and culinary purposes.

The introduction of gardening to Europe

In medieval times the monasteries were the main repositories of gardening knowledge and the important herbal lore. Though little is certainly known about the design and content of the monastic garden, it probably consisted of a walled courtyard, built around a well or arbour, with colour provided by flowers (some of which, including roses and lilies, served as ecclesiastical symbols), all of which maintained the ancient idea of the garden as a place of contemplation.

The earliest account of gardening in English, *The Feate of Gardening,* dating from about 1400, mentions the use of more than 100 plants, with instructions on sowing, planting, and grafting of trees and advice on cultivation of herbs such as parsley, sage, fennel, thyme, camomile, and saffron. The vegetables mentioned include turnip, spinach, leek, lettuce, and garlic.

Early gardening was largely for utility. The emergence of the garden as a form of creative display properly began in the 16th century. The Renaissance, with its increased prosperity, brought an upsurge of curiosity about the natural world and, incidentally, stirred interest in composing harmonious forms in the garden.

This awakening took especially firm root in Elizabethan England, which notably developed the idea that gardens were for enjoyment and delight. Echoing the Renaissance outlook, the mood of the period was one of exuberance in gardening, seen in the somewhat playful arrangements of Tudor times, with mazes, painted statuary, and knot gardens (consisting of beds in which various types of plants were separated by dwarf hedges). Flowers began to appear profusely in paintings and, as mentioned above, were used by poets in their verbal images.

This enthusiasm was accompanied by an earnest search for knowledge, and the period saw the birth of botanical science. A leading figure in this work was Carolus Clusius (Charles de l'Écluse), whose botanical skills and introduction of the tulip and other bulbous plants to the botanical gardens at Leiden, Neth., laid the foundation for Dutch prominence in international horticulture. The earliest botanical garden was that of Pisa (1543), followed by that of Padua (1545). The first in England was founded at Oxford in 1621, followed by Scotland's first, at Edinburgh, in 1667. The gardens at Kew, near London, were founded almost a century later, in 1759. These centres of experiment and learning have contributed greatly to the art and science of horticulture.

Early botanical gardens

The advances from the simple medieval style were marked and rapid at this time. The English statesman and scholar Francis Bacon could already, by 1625, advance a sophisticated and almost modern conception of the garden in his essay "On Gardens." He saw it as a place that should be planted for year-round enjoyment, offering a wide range of experiences through colour, form and scent, exercise and repose. The flower garden, already well established by the early 17th century, was set against a background of tall, clipped hedges and neatly scythed lawns. The taste of the time, as contemporary lists show, was for perfumed varieties such as carnations, lavender, sweet marjoram, musk roses, and poppies.

The plant trade. As interest in gardening developed in Europe the new trade of nurseryman was established, and the trade became highly important to the spread of knowledge and materials. By the end of the 17th century, nurserymen were relatively numerous in England, France, and the Low Countries, with keen customers among the nobility and gentry for all the exotica they could provide. The catalog of the Tradescant family's private botanical garden in London listed 1,600 plants in 1656. A number of them had been brought back by the family from visits to Virginia. These early exotica from the New World included now familiar plants such as the Michaelmas daisy, the Virginia creeper, hamamelis, goldenrod, the first perennial lupine, and such fine autumn-colouring trees as liquidambar and the staghorn sumac. The work of the nurserymen thus spread new plants more widely and, as breeding skills developed, contributed to the acclimatizing of foreign imports.

Vegetables and fruits. The history of vegetables is imprecise. Though familiar types, including the radish, turnip, and onion, are known to have been in cultivation from early times, it is fairly supposed that they were meagre and would bear little clear resemblance to modern equivalents. The early range available to European gardens and, later, to those in America, included such native plants as kale, parsnips, and the Brussels sprout family, with peas and broad beans grown as field crops.

Introduced and imported plants

The Romans introduced the globe artichoke, leek, cucumber, cabbage, asparagus, and the Mediterranean strain of garlic to their imperial territory, wherever these plants would flourish. Among plants imported to Europe from the Americas were the scarlet runner bean and tomato (both originally grown for ornament), corn (maize), and the vastly important potato. The numerous herbs in use were mostly native to European locations. One curiosity to the modern mind is that certain flowers, such as marigolds, violets, and primroses, were used as flavourings in the kitchen.

The cultivation of fruit trees was one of the most advanced skills and interests from the 16th century onward. Pride was taken in variety while, judging by the opulent still-life paintings of the period, the quality was remarkably high. Among the challenges bravely taken up in the 17th century in northern Europe was the growing of orange and lemon trees, though this was done more for the pleasure of their evergreen qualities than for their fruit. The catalog of the British royal gardens in 1708 shows 14 varieties of cherry, 14 apricots, 58 kinds of peach and nectarine, 33 plums, eight figs, 23 vines, 29 pears, and numerous varieties of apple.

The French style. The most favoured style for great house gardens in Europe during much of this period derived from the influence of the French designer André Le Nôtre, creator of the gardens at Versailles. The French style represented an extreme of formality, with box-edged parterres (elaborate and geometrical beds) typically placed near the residence to provide an arranged view. Trees were grouped in neat plantations or in bold lines along avenues, with terraces and statuary carefully placed to emphasize the architectural symmetry of the grand manner. The widespread adoption of this style among the European nobility and gentry reflected the potency of French cultural influence at the time. It was also related, on a practical basis, to the limited availability of planting materials, especially those offering autumn and winter display.

The change to a more natural style of gardening came about when, in the latter part of the 18th century, the opinion arose among leading gardeners, particularly those of the English gentry, that the formal manner brought with it a certain monotony. The increasing importation of foreign plants also brought with it opportunities for a large-scale transformation.

The plant hunters. The early importation of plants to Europe was managed through informal channels, following the increase in exploration and the spread of empires. Seeds and tubers were sent home by diplomats and missionaries, sea captains and travelers. An example of this type of collecting is afforded by Henry Compton, bishop of London, whose diocese included the American colonies. He was an avid collector, and he corresponded with like-minded experts in Europe and America and thus brought numerous fine plants to his exceptional garden in Fulham, west London. He also encouraged his missionaries to send home seeds. From one such source in Virginia came the *Magnolia virginiana,* the first magnolia to be cultivated. This was the beginning of what became known as the American garden, based upon magnolias, azaleas, and other woodland species.

As the appetite for exotica developed, plant collecting around the world became more systematized. Expeditions to foreign parts were organized and financed by nurserymen, botanical gardens, or syndicates of private gardeners. The botanist plant hunters thus sent out were exceptional and patient. They were required to endure long voyages and residence for up to several years in an often hostile environment. Their goal was to find the plant in flower, return in due season to collect seed, then see their delicate specimens back to Europe through varying climatic zones.

Collecting in North America and South America

North America's potential to yield countless new specimens was recognized early: the first book on American plants, published in London in 1577, was entitled *Joyfull Newes out of the New Founde Worlde* and was in itself a hint of the excited spirit of contemporary gardening. The jacaranda, flowering catalpa, and wisteria were among the finds made by Compton's missionaries in the Carolinas. An early resident collector in North America was John Bartram, regarded as the founder of American botany. He settled on a farm near Philadelphia in 1728 and, in 30 years of collecting in the Alleghenies, Carolinas, and other areas of North America, sent some 200 important plants to British gardens in sufficient quantity that they became widespread there.

The extremely rich west coast of North America was not exploited by plant collectors until the early 19th century. The contemporary importance of such discoveries is suggested by the fact that, in their celebrated crossing of the American continent in 1804–06, Lewis and Clark found time to collect the seeds of *Mahonia aquifolium* and *Symphoricarpos racemosus.* Perhaps the most distinguished collector among an exceptional fraternity was David Douglas, one of the numerous Scotsmen who contributed to international botany. His expeditions to the North American Far West brought to Europe such important timber trees as the Douglas fir, the Sitka spruce, the Monterey pine, and a number of now familiar shrubs such as *Garrya elliptica* and *Ribes sanguineum.* The California annuals he discovered made a lasting impact on the colour of Western gardens. In the 19th century plant collectors began to explore South America, where two Cornish brothers, William and Thomas Lobb, gained prominence. They are credited with carrying back to Europe the monkey puzzle tree (*Araucaria araucana*), native to the Andes mountains; the *Berberis darwinii;* and the *Escallonia macrantha.*

Collecting in China

Collectors went to a number of countries in the 19th century, but the most important area was China. Its flora was more intact than that in the West, because the erosions of the Ice Age had been less severe for climatic reasons, and it had a long history of skilled gardening. Plant collection was difficult, however, because for many years the only foreigners allowed to travel within its borders were Jesuit priests. They aided botanists by sending many specimens to Paris and London. The first professional collector to live in China was William Kerr, who sent out 238 new plants. Real exploration of the interior did not begin until the 1840s. China, Japan, and the Himalayas produced unparalleled riches in rhododendrons, azaleas, flowering cherries, ornamental maples, roses, lilies, primulas, poppies, kerrias, and quinces.

The conditions for transporting plants from such distances had been much improved by Nathaniel B. Ward's invention of the wardian case, an airtight glass box that protected the plants from sea air and harsh climate. Gradually almost all regions and countries were visited, and new plants and their progeny were dispersed around the Western world. And still the search for new specimens continues.

From the 19th century. By the early 19th century, with the expansion of the horticultural trade, gardening had become international in scope. Numerous handbooks spread knowledge. The founding of new garden and botanical societies, such as the London (later Royal) Horticultural Society, helped to increase interest, encourage science, and raise standards. Such moves signaled the rise of the small leisure gardener; a floral retreat was no longer the sole property of the rich. It now extended from the manor to the small suburban garden.

Gardens in North America had generally been smaller and trimmer than their European counterparts, with box edgings and pleached trees (that is, lines of trees allowed to grow with branches interlaced to form a screen) as seen in the reconstructed gardens of Williamsburg, Va. The "natural" gardening style (known on the European continent as the English style), which had overtaken earlier formality, allowed wider use of plant varieties. This approach became the pervasive trend in the west, notably through the views of John Claudius Loudon, whose *Encyclopaedia of Gardening* (1822) set the pattern of domestic cultivation over a long period with a style known as Gardenesque. His style encouraged the individual qualities of garden elements while ensuring that together they made a harmonious blend.

The "natural" style

The natural style was further enhanced by an English artist and landscape architect, Gertrude Jekyll. In her opinion, the first purpose of a garden is to give happiness and repose of mind. With experience derived from the richly floral cottage gardens of Surrey, she developed the idea of supporting plants with an architectural base and allowing them to grow in a free form, encouraging natural shape and creating harmonious relationships of colour.

The period saw much progress in garden equipment and supplies. Heated greenhouses had been in use since the late 17th century, and mass production led to great strides in nursery gardening. The modern, bladed lawn mower was first seen in a design of 1832; in more recent times the application of the jet-engine principle led to the hover mower. Fertilizer development was also important, from the discovery of superphosphate to the devising of modern kinds of foliar feeding.

In the second half of the 20th century interest in gardening brought in new adherents in unprecedented numbers; they were advised and encouraged by numerous publications and by television and radio programs. Though the process was very gradual, domestic gardening became somewhat more adventurous. Among the more ambitious, designs took a multiplicity of forms, from the Japanese garden, producing an austere magic out of rock and pebble, to the other extreme of the wild country garden, virtually left to seed itself. Increasing numbers of professional designers at their best set high standards to emulate. But the art of gardening still depends on a simple empathy with the needs and nature of living things. Symbolic of this essential, the spade has remained much the same implement that it had been in medieval times.

TYPES OF GARDENS

The domestic garden can assume almost any identity the owner wishes within the limits of climate, materials, and means. The size of the plot is one of the main factors, deciding not only the scope but also the kind of display and usage. Limits on space near urban centres, as well as the wish to spend less time on upkeep, have tended to make modern gardens ever smaller. Paradoxically, this happens at a time when the variety of plants and hybrids has never been wider. The wise small gardener avoids the temptations of this banquet. Some of the most attractive miniature schemes, such as those seen in Japan or in some Western patio gardens, are effectively based on an austere simplicity of design and content, with a handful of plants given room to find their proper identities.

In the medium- to large-sized garden the tradition generally continues of dividing the area to serve various purposes: a main ornamental section to enhance the residence and provide vistas; walkways and seating areas for recreation; a vegetable plot; a children's play area; and features to catch the eye here and there. Because most gardens are mixed, the resulting style is a matter of emphasis rather than exclusive concentration on one aspect. It may be useful to review briefly the main garden types.

Flower gardens. Though flower gardens in different countries may vary in the types of plants that are grown, the basic planning and principles are nearly the same, whether the gardens are formal or informal. Trees and shrubs are the mainstay of a well-designed flower garden. These permanent features are usually planned first, and the spaces for herbaceous plants, annuals, and bulbs are arranged around them. The range of flowering trees and shrubs is enormous. It is important, however, that such plants be appropriate to the areas they will occupy when mature. Thus it is of little use to plant a forest tree that will grow 100 feet (30 metres) high and 50 feet across in a small suburban front garden 30 feet square, but a narrow flowering cherry or redbud tree would be quite suitable.

Blending and contrast of colour as well as of forms are important aspects to consider in planning a garden. The older type of herbaceous border was designed to give a maximum display of colour in summer, but many gardeners now prefer to have flowers during the early spring as well, at the expense of some bare patches later. This is often done by planting early-flowering bulbs in groups toward the front. Mixed borders of flowering shrubs combined with herbaceous plants are also popular and do not require quite so much maintenance as the completely herbaceous border.

The old-fashioned garden

Groups of half-hardy annuals, which can withstand low night temperatures, may be planted at the end of spring to fill gaps left by the spring-flowering bulbs. The perpetual-flowering roses and some of the larger shrub roses look good toward the back of such a border, but the hybrid tea roses and the floribunda and polyantha roses are usually grown in separate rose beds or in a rose garden by themselves.

Woodland gardens. The informal woodland garden is the natural descendant of the shrubby "wilderness" of earlier times. The essence of the woodland garden is informality and naturalness. Paths curve rather than run straight and are of mulch or grass rather than pavement. Trees are thinned to allow enough light, particularly in the glades, but irregular groups may be left, and any mature tree of character can be a focal point. Plants are chosen largely from those that are woodlanders in their native countries: rhododendron, magnolia, pieris, and maple among the trees and shrubs; lily, daffodil, and snowdrop among the bulbs; primrose, hellebore, St.-John's-wort, epimedium, and many others among the herbs.

Rock gardens. Rock gardens are designed to look as if they are a natural part of a rocky hillside or slope. If rocks are added, they are generally laid on their larger edges, as in natural strata. A few large boulders usually look better than a number of small rocks. In a well-designed rock garden, rocks are arranged so that there are various exposures for sun-tolerant plants such as rockroses and for shade-tolerant plants such as primulas, which often do better in a cool, north-facing aspect. Many smaller perennial plants are available for filling spaces in vertical cracks among the rock faces.

The main rocks from which rock gardens are constructed are sandstone and limestone. Sandstone, less irregular and pitted generally, looks more restful and natural, but certain plants, notably most of the dianthuses, do best in limestone. Granite is generally regarded as too hard and unsuitable for the rock garden because it weathers very slowly.

Water gardens. The water garden represents one of the oldest forms of gardening. Egyptian records and pictures of cultivated water lilies date as far back as 2000 BC. The Japanese have also made water gardens to their own particular and beautiful patterns for many centuries. Many have an ornamental lantern of stone in the centre or perhaps a flat trellis roof of wisteria extending over the water. In Europe and North America, water gardens range from formal pools with rectangular or circular outline, sometimes with fountains in the centre and often without plants or with just one or two water lilies (*Nymphaea*), to informal pools of irregular outline planted with water lilies and other water plants and surrounded by boggy or damp soil where moisture-tolerant plants can be grown. The pool must contain suitable oxygenating plants to keep the water clear and support any introduced fish. Most water plants, including even the large water lilies, do well in still water two to five feet deep. Temperate water lilies flower all day, but many of the tropical and subtropical ones open their flowers only in the evening.

In temperate countries water gardens also can be made under glass, and the pools can be kept heated. In such cases, more tropical plants, such as the great *Victoria amazonica* (*V. regia*) or the lotus (*Nelumbo nucifera*), can be grown together with papyrus reeds at the edge. The range of moisture-loving plants for damp places at the edge of the pool is great and includes many beautiful plants such as the candelabra primulas, calthas, irises, and osmunda ferns.

Herb and vegetable gardens. Most of the medieval gardens and the first botanical gardens were largely herb gardens containing plants used for medicinal purposes or herbs such as thyme, parsley, rosemary, fennel, marjoram, and dill for savouring foods. The term herb garden is usually used now to denote a garden of herbs used for cooking, and the medicinal aspect is rarely considered. Herb gardens need a sunny position because the majority of the plants grown are native to warm, dry regions.

The vegetable garden also requires an open and sunny location. Good cultivation and preparation of the ground are important for successful vegetable growing, and it is also desirable to practice a rotation of crops as in farming. The usual period of rotation for vegetables is three years; this also helps to prevent the carryover from season to season of certain pests and diseases.

The French *potager*

The old French *potager*, the prized vegetable garden, was grown to be decorative as well as useful; the short rows with little hedges around and the high standard of cultivation represent a model of the art of vegetable growing. The elaborate parterre vegetable garden at the Château de Villandry is perhaps the finest example in Europe of a decorative vegetable garden.

Specialty gardens. *Roof gardens.* The modern tendency in architecture for flat roofs has made possible the development of attractive roof gardens in urban areas above private houses and commercial buildings. These gardens follow the same principles as others except that the depth of soil is less, to keep the weight on the rooftop low, and therefore the size of plants is limited. The plants are generally set in tubs or other containers, but elaborate roof gardens have been made with small pools and beds. Beds of flowering plants are suitable, among which may be stood tubs of specimen plants to produce a desired effect.

Scented gardens. Scent is one of the qualities that many people appreciate highly in gardens. Scented gardens, in which scent from leaves or flowers is the main criterion for inclusion of a plant, have been established, especially for the benefit of blind people. Some plants release a strong scent in full sunlight, and many must be bruised or

(Left) T.A. Fretz—Color Advantage; (right) Pamela J. Harper

(Left) A border of perennial plants in an English flower garden. (Right) Garden in California where a variety of succulents are being grown.

rubbed to yield their fragrance. These are usually grown in raised beds within easy reach of visitors.

CONTENTS OF GARDENS

Permanent elements. The more or less permanent plants available for any garden plan are various grasses for lawns, other ground-cover plants, shrubs, climbers, and trees. More transitory and therefore in need of continued attention are the herbaceous plants, such as the short-lived annuals and biennials, and the perennials and bulbous plants, which resume growth each year.

Lawns and ground covers. Areas of lawn, or turf, provide the green expanse that links all other garden plantings together. The main grasses used in cool areas for fine-textured lawns are fescues (*Festuca* species), bluegrasses (*Poa* species), and bent grasses (*Agrostis* species), often in mixtures. A rougher lawn mixture may contain ryegrass (*Lolium* species). In drier and subtropical regions, Bermuda grass (*Cynodon dactylon*) is frequently used, but it does not make nearly as fine a lawn as those seen in temperate regions of higher rainfall.

Ground covers are perennial plants used as grass substitutes in regions where grasses do poorly, or they are sometimes combined with grassy areas to produce a desired design. The deep greens, bronzes, and other colours that ground-cover plants can provide offer pleasing contrasts to the green of a turf. Ground covers, however, are not so durable as lawns and do not sustain themselves as well under foot traffic and other activities. Among the better known plants used as ground covers are Japanese spurge (*Pachysandra terminalis*), common periwinkle (*Vinca minor*), lily of the valley (*Convallaria majalis*), ajuga, or bugleweed (*Ajuga reptans*), many stonecrops (*Sedum* species), dichondra (*Dichondra repens*), and many ivies (*Hedera* species).

Shrubs and vines (climbers). Smaller woody plants, such as shrubs and bushes, have several stems arising from the base. These plants attain heights up to about 20 feet. They often form the largest part of modern gardens because their cultivation requires less labour than that of herbaceous plants, and some flowering shrubs have extended blooming periods. Among the popular garden shrubs are lilac (*Syringa vulgaris*), privet (*Ligustrum* species), spirea (*Spiraea* species), honeysuckle (*Lonicera* species), forsythia (*Forsythia* species), mock orange (*Philadelphus* species), and hydrangea (*Hydrangea* species).

Bushlike azaleas and rhododendrons (both of which are species of *Rhododendron*) provide colourful blossoms in spots where there is semishade.

Climbers are often useful in softening the sharp lines of buildings, fences, and other structures. They can provide shade as an awning or cover on an arbour or garden house. Some species are also useful as ground covers on steep slopes and terraces. Among the many woody perennial climbers for the garden are the ivies, trumpet creeper (*Bignonia*, or *Campsis, radicans*), clematis (*Clematis* species), wisteria (*Wisteria sinensis*), climbing roses, annual herbaceous vines such as morning glory (*Ipomoea* species), and ornamental gourds, the last of which can provide rapid but temporary coverage of unsightly objects.

Trees. Trees are the most permanent features of a garden plan. The range of tree sizes, shapes, and colours is vast enough to suit almost any gardening scheme, from shrubby dwarf trees to giant shade trees, from slow to rapid growers, from all tones of green to bronzes, reds, yellows, and purples. A balance between evergreen trees, such as pines and spruces, and deciduous trees, such as oaks, maples, and beeches, can provide protection and visual interest throughout the year.

Transitory elements. *Herbaceous plants.* Herbaceous plants, which die down annually and have no woody stem above ground, are readily divided into three categories, as mentioned earlier: (1) Annuals, plants that complete their life cycle in one year, are usually grown from seed sown in the spring either in the place they are to flower or in separate containers, from which they are subsequently moved into their final position. Annuals flower in summer and die down in winter after setting seed. Many brilliantly coloured ornamental plants as well as many weeds belong in this category. Examples of annuals are petunia and lobelia. (2) Biennials are plants sown from seed one year, generally during the summer. They flower the second season and then die. Examples are wallflower and sweet william. (3) Herbaceous perennials are those that die down to the ground each year but whose roots remain alive and send up new top growth each year. They are an important group in horticulture, whether grown as individual plants or in the assembly of the herbaceous border. Because they flower each year, they help to create the structure of the garden's appearance, so their placement must be considered carefully. Examples are delphinium and lupine.

Annuals, biennials, and herbaceous perennials

Bulbous plants. For horticultural purposes, bulbous plants are defined to include those plants that have true bulbs (such as the daffodil), those with corms (such as the crocus), and a few that have tubers or rhizomes (such as the dahlia or iris). A bulb is defined as a modified shoot with a disklike basal plate and above it a number of fleshy scales that take the place of leaves and contain foods such as starch, sugar, and some proteins. Each year a new stem arises from the centre. A corm consists of the swollen base of a stem, generally rounded or flattened at the top and covered with a membranous tunic in which reserve food materials are stored. A tuber or rhizome is not the base of the stem but rather a swollen part of an underground stem; it is often knobbly. All such plants have evolved in places where they can survive in a semidormant state over long unfavourable seasons, either cold, mountain winters or long, droughty summers.

THE PRINCIPLES OF GARDENING

Soil: its nature and needs. Soil is the basic element in the cultivation of all plants, although soilless growth in water with or without gravel or sand, enriched with suitable chemicals (hydroponics) can be very successful.

Soil consists of particles, mainly mineral, derived from the breakdown of rocks and other substances together with organic matter. In the pore spaces between the particles both water (containing dissolved salts) and air circulate. The air contains more carbon dioxide and less oxygen than does the atmosphere. Minute living organisms are also present in soil in immense quantities and are what make it "alive." Plants must penetrate this pore space to reach much of their nourishment.

Soil fertility

The soil must be managed for fertility (the ability to supply plant nutrients) and physical condition. Nutrients must be supplied and released in forms available to the plant. Sixteen elements are necessary for plant growth. Three of these, carbon, oxygen, and hydrogen, are provided through water and air; the other 13 are provided through the soil. The elements required in relatively large amounts are called major elements: nitrogen, phosphorus, potassium, calcium, magnesium, and sulfur. The minerals required in small quantities are called trace elements: iron, boron, manganese, zinc, molybdenum, copper, and chlorine.

Soils can be roughly divided into three main types on the basis of their usefulness horticulturally, but many areas contain a mixture.

Clays. Clays, in which the particles are very fine, are called in horticulture heavy soils, because it is difficult to turn them over with a spade. They can be very fertile but tend to be lacking in good drainage, holding their water closely adhered to the soil particles; therefore, they cannot be worked when wet, and under pressure they tend to compact tightly, driving out the air. During drought they tend to become hard and even to develop large cracks so that they cannot be worked satisfactorily. Clay soils can be lightened with as much humus as can be dug into them. Humus may be any decayed organic matter, such as farmyard manure, leaf mold, or compost made from kitchen scraps and grass clippings.

Sands and gravels. Sands and gravels are opposite in properties to clay. The soil particles are large, and the soils are called light because they are easy to work and turn in nearly all weather. Since their water-holding capacity is very low, however, they tend to dry out quickly. They are "hungry" soils requiring great quantities of manures, humus, and fertilizers to keep them prolific.

Peats and heaths. Peats and heaths are usually very acid and ill-drained. They result where conditions have prevented the complete breakdown of old vegetable matter into humus, generally because of poor aeration and surplus acid bog water. Much peat is derived from the decaying roots of sphagnum moss, useful for mulching in the garden. A heath soil is generally less fertile, consisting of a large mixture of sand with the peat and tending to be very low in mineral content and in water-retaining capacity.

The ideal garden soil is a medium loam consisting of a mixture of clay and sand, fairly rich in humus and easily worked, and not forming large clods when dry. The consistency of the soil is important, for a porous, properly tilled soil provides a medium through which roots can penetrate readily and rapidly. Another factor of importance in soils is the degree of acidity or alkalinity. Soil alkalinity is usually derived from free calcium carbonate or a similar alkaline salt. Soil reaction can be modified. It may be made more alkaline by adding one of the organic salts, of which calcium is best, in the form of lime. Acidity may be increased by adding hydrogen, in the form of sulfur compounds such as ammonia sulfate or superphosphate.

Feeding: fertilizing and watering. Maximum return can be obtained only from soil with an ample supply of elements necessary for plant growth, combined with sufficient moisture to enable them to be dissolved and absorbed through the plant hairs.

Treatment with farmyard manure or garden compost can supply the majority of these requirements. Because manure and compost are scarce in urban areas it is often necessary to use mineral fertilizers as well as organics. The soil is such a complex substance that all fertilizers must be applied in moderation and in balance with each other according to the deficiencies of the soil and the requirements of the particular crop. Different crops have different fertilizer needs. Manures are generally best dug into the ground in autumn in a temperate climate but also may be used as mulches in spring to control weeds. A mulch is a surface layer of organic matter that helps the several needs of feeding, conserving moisture, and controlling weeds. Black polyethylene sheeting is now widely used for all the mulching functions except feeding.

Watering

Watering of newly placed plants and of all plants during periods of drought is an essential gardening chore. Deep and thorough watering—not simply sprinkling the soil surface—can result in greatly improved growth. Water is essential in itself, but it also makes minerals available to plants in solution, the only form usable by plants. About one inch of water applied each week to the soil surface will percolate down about six inches; this is a minimal subsistence amount for many herbaceous garden plants, and small trees and shrubs require more. Proper watering once a week encourages deep penetration of roots, which in turn enables plants to survive dry surface conditions.

Drainage. Drainage is the other important side of water management. All plants need water but the amount needed varies, and if plants are forced to absorb more than they need, a form of drowning occurs. The symptoms are most easily seen in overwatered pot plants but are also visible to an experienced eye in badly drained corners of a garden. Roots require air as well as water and depend on subsurface water to bring the necessary oxygen. In large private gardens and in commercial gardens, buried earthenware piping is commonly used. In smaller gardens drainage can be readily achieved by the use of sumps, that is, holes dug to a depth of about four feet in affected places. The bottom half of the sump is filled with stones, through which excess water drains. Such measures may greatly improve the potential of a garden and the workability of its soil.

Protecting plants. Most plants have a precise level of tolerance to cold, below which they are killed. Many plants from tropical or subtropical regions cannot survive frost and are killed by temperatures below 32° F (0° C). These are called frost-tender. Others, called half-hardy, can withstand a few degrees of frost. Fortunately many of the best garden plants are completely hardy, a quality often encouraged by careful breeding, and will withstand any low temperatures likely to be reached in temperate regions.

Various measures can be taken to give frost protection, from the simple ones appropriate for smaller gardens to the elaborate coverings used to protect valuable horticultural crops. Removing weeds that shade the soil increases the amount of heat stored during the day. Well-drained soil is less susceptible. Any shield against wind in frosty weather enhances survival capability. The simplest form of protection is a wrapping to keep warmer air around the plant. This can be a mulch (leaves, soil, ashes) placed over the crown of a slightly tender plant in winter or a shield of sacking for leaf-shedding plants (not as desirable for evergreens, which utilize their leaves all the year).

Forcing flowers

Glass structures such as greenhouses or outdoor frames can provide additional protection for tender plants. Such structures can be heated and the temperature regulated by a thermostat to any required degree. Thus in temperate regions, orchids and other tropical plants can be grown so that they flower throughout the winter, many being forced to flower earlier than their normal season by the higher temperature. Greenhouses are divided by gardeners into four rough categories: (1) The cold house, in which there is no supplementary heating and which is suitable only for plants that will not be killed by a few degrees of frost (such as alpines or potted bulbous plants). The combination of heat from the sun and protection from wind will keep such a house appreciably warmer than the temperature outside. (2) The coolhouse, in which the minimum temperature is kept to 45° F (7° C). Most amateurs' greenhouses fall into this class, and a very large range of plants can be grown in them. (3) The intermediate house, in which the minimum temperature is kept at 55° or 60° F (13° or 16° C), and which is suitable for a wide range of orchids. (4) The hothouse, or stove house, in which the minimum temperature is kept above 60° F (16° C) and in which tropical plants such as anthuriums and cattleyas (a genus of the orchid family) can be grown.

Training and pruning. Training, the orienting of the plant in space, is achieved by techniques that direct the shape, size, and direction of plant growth. It may be accomplished by use of supports to which plants can be bent, twisted, or fastened. Pruning, the judicious cutting away of plant parts, is performed for other purposes: to contain size, to encourage fruiting in orchard trees, or to improve the appearance of ornamental trees and shrubs. It is one of the most important horticultural arts.

Where trees and shrubs are left to grow naturally, they often become much too large for their space in the garden. Also they may grow lanky and misshapen and have much dead growth. Where a branch or shoot is cut, it will often be induced to make a number of young shoots from below the cut, and these are likely to flower more freely than the older branches. Fruit trees in particular when pruned annually often give fruit of finer quality, larger in size, freer of disease, and of better colour. The two basic pruning cuts are known as heading back and thinning out. Heading back consists of cutting back the terminal portion of a branch to a bud; thinning out is the complete removal of a branch to a lateral or main trunk. Heading back, usually followed by the stimulation of lateral budbreak below the cut, produces a bushy, compact plant, suitable for a hedgerow, and it is often used to rejuvenate shrubs that have become too large or that flower poorly. Thinning out, which encourages longer growth of the remaining terminals by reducing lateral branches, tends to open up the plant, producing a longer plant. In general, pruning, started when the plant is young, obviates the need for drastic and risky remedial pruning later of a large, old, or misshapen bush or tree.

Pruning systems

Particular spatial arrangements may increase light utilization, facilitate harvesting or disease control, or improve productivity and quality. Thus, training and pruning form an essential part of fruit growing throughout the life of the plant. Special attention is given in the formative years to obtain desired shape and structure. The key to training is the point on the main stem from which branches form. In the central-leader system of training, the trunk forms a central axis with branches distributed laterally up and down and around the stem. In the open-centre or vase system the main stem is terminated and growth forced through a number of branches originating close to the upper end of the trunk. An intermediate system is called the modified-leader system. In espalier systems, plants are trained to grow flat along a wire or trellis (see Figure 1). Properly executed espaliers are extremely attractive as ornamentals. Espaliers in combination with dwarfing rootstocks allow high density orchards that are very productive on a per-unit-area basis with the fruit close to the ground for easy harvest. Extensive pruning is required annually to maintain the system.

There are a number of physiological responses to training and pruning. Orientation of the plant may have a marked effect on growth and fruiting. Thus fruit trees planted on an inclined angle become dwarfed and flower earlier; training branches in a horizontal position produces the same effect. This effect is achieved naturally when a heavy fruit load bends a limb down. The main effects of pruning are achieved by altering the root–shoot balance. Thus an explosion of vegetative growth normally occurs after extensive shoot pruning. Severely pruned plants, especially if they are in the juvenile stage of growth, tend to remain vegetative. Similarly the slowdown of vegetative growth by root pruning encourages flowering.

The training of plants to grow in unnatural shapes for ornamental purposes is called topiary. In Roman and Renaissance times, when ingenious topiary was in high fashion, plants were trained to unusual and fantastic shapes such as beasts, ships, and building facades. Though more modestly, hedges and shrubs are still trained to geometric shapes in formal gardens.

Another extreme form of training is the Japanese art of bonsai, the creation of dwarfed potted trees by a combination of pruning (both roots and tops) and restricted nutrition. Living trees more than 100 years old and only a few feet high are grown in special containers arranged to resemble the natural landscape.

By courtesy of (B, C) Henry Leuthardt Nurseries, from (A) *Arnoldia* (Aug 29, 1969)

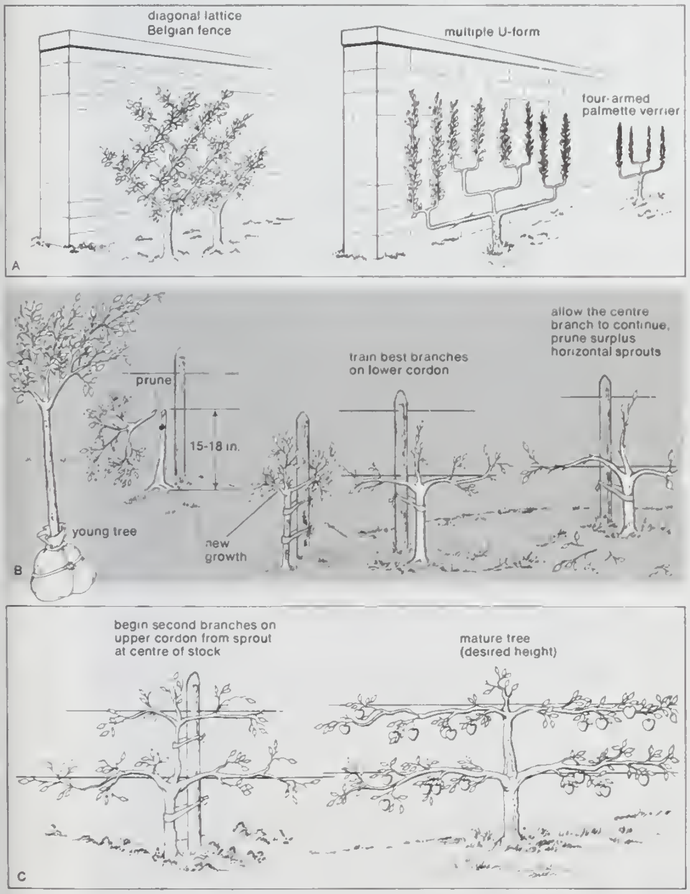


Figure 1: *The espalier technique for training fruit trees and shrubs.*
(A) Examples of espalier patterns. (B) Planting and training an espalier. (C) Final shaping and (right) tree in its finished form, a double horizontal cordon.

Propagation. New plants are produced either from seed or by the techniques of division, taking cuttings, grafting, budding, or layering. (A fuller description of propagation and breeding processes is found below in the section *Horticulture.*) For the ordinary gardener propagation is a relatively simple but interesting process normally used for economic provision of more versions of favourite plants; as part of exchanges with other gardeners; or as a wise precaution against winter losses.

Propagation by cuttings is the most common practice. Young shoots of the current season are usually the most successful at rooting. Roses are usually propagated by budding, in which a bud from the rose desired is inserted in rootstock (that part of the plant tissue from which a root can form) just above ground level. Fruit trees are usually propagated by layering, in which a young shoot is pegged down in the ground with the end twisted upward almost at right angles; the lower side of the wood just before the twist is wounded so as to induce rooting. When this has taken place, the layer is severed from the parent.

Control of weeds. Controlling weeds is a basic, and probably the most arduous, factor of cultivation and has been carried on from the time the earliest nomads settled down to an agricultural life. It has always been necessary to free the chosen crops of competition from other plants. For smaller weeds hoeing is practicable. The weeds are cut off by the action of the hoe and left to wither on the surface. Hand weeding, by pulling out individual weeds, is often necessary in gardens, particularly the rock garden, in seed boxes, and in the herbaceous border or among annuals. Chemical and biological control of weeds developed greatly after World War II and has made much mechanical cultivation unnecessary.

Control of pests and diseases. Damage to plants is most often caused by pests such as insects, mites, eelworms, and other small creatures but may also be caused by mammals such as deer, rabbits, and mice. Damage by disease is that caused by fungi, bacteria, and viruses.

Pest control through vigilance

Prevention is generally better than cure, and constant vigilance is necessary to prevent a pest infestation or a disease outbreak. Control can be obtained by the use of chemical sprays, dusts, and fumigants, but some of these are so potent that they should be used only by the experienced operator. Considerable evidence is available regarding the possible harmful long-term effects on the biological chain of excessive use of some of these noxious chemicals, particularly the hydrocarbons. Some control can be obtained through good garden practices: clearing up all dead and diseased material and burning it; pruning and thinning so that a reasonable circulation of air is obtained through the plants; and crop rotation. Some control may also be obtained through the natural biological predators. The breeding of plants immune to certain pests and diseases is also a valuable means of control.

Mechanical aids. Mechanical devices to aid the gardener include tillers, lawn mowers, hedge cutters, sprinklers, and a variety of more esoteric equipment that has made gardening an easier pursuit. Such machines are not a substitute for good judgment and technique in the garden, however, nor will they give anyone a completely labour-free garden. They do enable a considerably larger area to be cultivated and maintained than if all labour is performed by hand. (G.A.C.H./P.M.Sy./Ro.P.)

Raising houseplants

The range of houseplants

Many exotic plants native to warm, frost-free parts of the world can be grown indoors in colder climates, in portable containers or miniature gardens. Most houseplants are derived from plants native to the tropics and near tropics. Those that make the best indoor subjects are the species that adjust comfortably to the rather warm, dry conditions that generally prevail in indoor living spaces.

Although many plants can be grown successfully indoors, there are certain groups that, because of their attractiveness and relative ease of maintenance, are generally considered the best houseplants. These include the aroids, bromeliads, succulents (including cacti), ferns, begonias, and palms, all of which have long been favourites. Somewhat more

Selection of houseplants that, with care, grow well indoors.
Pamela J Harper

demanding are those that are grown primarily for their flowers—African violets, camellias, gardenias, geraniums (*Pelargonium* species), and orchids.

HISTORICAL BACKGROUND

Paintings and sculptures make clear that the practice of indoor gardening can be traced at least to the early Greeks and Romans, who grew plants in pots and perhaps brought them into their homes. The older civilizations of Egypt, India, and China also made use of pot plants but usually in outdoor situations, often in courtyards that were extensions of the house; and for centuries the Japanese have carried on the dwarfing of trees and other plants for room ornaments. But the popular art of growing houseplants did not receive much comment until the 17th century, when, in *The Garden of Eden* (1652), Sir Hugh Platt, an English agricultural authority, wrote of the possibility of cultivating plants indoors. Shortly thereafter, glasshouses (greenhouses) and conservatories, which had been used during Roman times to force plants to flower, were built in England and elsewhere to house exotic plants. In mid-19th-century England and France, books began to appear on the growing of plants in private residences, and the use of enclosed glass cases of plants (the wardian cases, or terraria) became popular.

TYPES OF HOUSEPLANTS

There are thousands of tropical and subtropical plants that can adapt to growing indoors. Although some fancy exotic species do well only in a humid conservatory or a glass-enclosed terrarium, a great many species have been introduced that endure the adverse conditions of dry heat and low light intensity that prevail in many houses. A selection of the more widely favoured houseplants follows, under two sections: foliage plants, some of which also bear interesting flowers; and flowering plants, species kept primarily for their flowers.

Foliage plants. In the aroid family, which has provided a range of long-lived houseplants, most prominent are the philodendrons. These are handsome tropical American plants, generally climbers, with attractive leathery leaves, heart-shaped, and often cut into lobes. *Monstera deliciosa,* or *Philodendron pertusum,* the Swiss cheese plant, has showy, glossy, perforated leaves slashed to the margins.

The dumb canes, of the genus *Dieffenbachia,* appear in a number of attractive species. They are handsome tropical foliage plants usually with variegated leaves; they tolerate neglect and thrive even in dry rooms. The Chinese evergreens, of the genus *Aglaonema,* are fleshy tropical Asian herbs of slow growth, with leathery leaves often bearing silvery or colourful patterns; they are durable and are tolerant of indoor conditions. Members of *Scindapsus,* popularly known as pothos, or ivy-arums, are tropical climbers from the Malaysian monsoon area; their variegated leaves are usually small in the juvenile stage. They do well in warm and even overheated rooms. The peace lilies (not a true lily), of the genus *Spathiphylla,* are easy-growing, vigorous tropical herbs forming clumps; they have green foliage and a succession of flowerlike leaves (spathes), usually white. Species of *Anthurium,* many of which, such as the flamingo flower, have colourful spathes, do best in humid conditions. *Caladium*'s tropical American tuberous herbs produce fragile-looking but colourful foliage; they keep surprisingly well if protected from chills and wintry drafts.

Begonias, with their often very decorative leaves, have long been favourites among houseplants, but, with few exceptions, they require more humidity and fresh air than the modern home provides. *Begonia metallica,* with its olive-green, silver-haired foliage; *B. masoniana,* with beautiful green, puckered leaves splotched brown; and *B. serratipetala,* with small leaves spotted pink, are examples of types more resistant to dry rooms.

Small foliage plants

There are many small foliage plants, often with strikingly patterned foliage, native to the tropical forest floor, some of which have become remarkably good houseplants. Among them are several prayer plants (*Maranta* species), which fold their attractive leaves at night; and the exquisite *Calathea makoyana,* or peacock plant, with translucent foliage marked with a feathery peacock design. *Pilea cadierei,* or aluminum plant, is easy to grow; it has fleshy leaves splashed with silver. *Codiaeum* species, or crotons, are multicoloured foliage plants that need maximum light and warmth to hold their leaves and coloration well. Although primarily thought of as bedding plants, the varicoloured coleuses, or painted nettles, can decorate a sunny window with a brilliant array of leaf patterns. *Peperomia* species form miniature rosettes or vines with waxy foliage, corrugated and decorated either with silver or creamy white.

Bromeliads constitute a plant family peculiar to the Western Hemisphere; they dwell on trees and rocks (as epiphytic plants) or on the forest floor (as terrestrial plants) and usually form rosettes of leathery, concave leaves, many with bizarre designs or striking variegations. Their flowers may be hidden deep in the centre of the rosette, surrounded by a cup of brilliant crimson inner leaves, as in *Neoregelia* and *Nidularium.* Species of *Aechmea* and *Guzmania* form colourful spikes or heads of long-lasting leathery bracts or bright berries. *Billbergia* species are tubular in shape; their showy flower stalk, with blue flowers, is often pendant. Most forms of *Tillandsia* and *Vriesea* have spear-shaped, flattened, colourful flower spikes. The earth stars of the terrestrial genus *Cryptanthus* are more or less flattened rosettes with striking leaf design, mottled, striped, or tiger-banded in silver over greens and bronzes.

Succulents. Cacti, most members of which are native to the Western Hemisphere, have developed a special capacity to store water in thick, fleshy bodies. They thrive in much sunlight and need very little water. There are many often curious forms: the tiny button cactus, *Epithelantha;* the myriad pincushion species of *Mammillaria; Parodia,* or Tom Thumb cactus; and *Rebutia,* the pygmy cactus. The last two bloom when young and tiny. Other forms include *Gymnocalycium,* or chin cactus; *Notocactus,* or ball cactus; *Echinocactus,* known as barrel cactus; various *Opuntia* species, including bunny ears and chollas; and *Cephalocereus,* or old-man cactus, with its glistening white hair. Larger cacti include *Cereus* and its relatives, often night-blooming, and the giants of the desert, such as the saguaro (classified as *Cereus giganteus* or as *Carnegiea gigantea*), with branching columns up to 50 feet (15 metres) in height. Cacti of tropical forests include the epiphytic *Rhipsalis,* found also in Africa, Madagascar, and Sri Lanka, and the near-epiphytic leaf, or orchid, cacti, *Epiphyllum,* which bloom in many colours.

Succulents other than cacti have also contributed favourite subjects for indoor growing. A typical stem succulent is *Euphorbia,* with its often angled candelabra-like columns resembling those of cacti. Leaf succulents are represented by *Aloe,* famous since ancient times as a medicinal plant; *Echeveria,* or hen and chickens; *Kalanchoe tomentosa,* the panda plant; *Crassula,* the jade plant; and *Haworthia,* which has rosettes with pearly dotted leaves. Durable pot plants include the strap-leaf snake plants, or *Sansevieria* species; they are remarkable for tolerating much neglect and growing in less than ideal locations.

Trees. *Dracaena,* the dragon trees, includes such houseplants as *D. marginata,* from Madagascar, which forms clusters of twisted stems topped by rosettes of narrow, leathery leaves. Other examples are *D. deremensis* 'Warneckei,' with its handsome, symmetrical rosette of sword-shaped, milky-green leaves with white stripes; and *D. sanderiana,* the ribbon plant, a diminutive and slender, highly variegated species that can be grown in water. Similar in appearance is *Pandanus veitchii,* which has a rosette of leathery, sword-shaped leaves—glossy green and banded white—arranged in spirals.

Sub-tropical evergreens

Several subtropical evergreens can be grown in cooler locations indoors. Preeminent among them is the Norfolk Island pine (*Araucaria heterophylla,* or *A. excelsa*)—not a true pine—an undemanding graceful conifer with tiered branches of fresh green needles; it is long-lived even in dim corners in any temperature above freezing. *Podocarpus,* the somber Buddhist pine, forms dense pyramids of dark-green needlelike leaves; it also prefers cooler locations.

(Top left, bottom left, centre, and right) Barry L. Runk—Grant Heilman; (top right) Runk/Schoenberger—Grant Heilman, (centre right) George Taloumis

Common houseplants.
(Top left) African violets (*Saintpaulia*); (top right) spider plant (*Chlorophytum comosum*); (centre right) rabbit's foot fern (*Davallia fejeensis*); (bottom left) weeping fig (*Ficus benjamina*); (bottom centre) peperomia (*Peperomia obtusifolia*); (bottom right) jade plant (*Crassula argentea*).

Among the many broad-leaved woody evergreens used as houseplants is *Brassaia actinophylla,* the umbrella tree, better known as *Schefflera.* Its spreading crowns of palmately divided, glossy green leaves do best in a light and warm location. Another picturesque plant is *Polyscias fruticosa,* the Ming aralia, with willowy, twisting stems densely clothed toward their tops with fernlike, lacy foliage.

The so-called rubber trees of the genus *Ficus* are widely used in homes and offices. All require good light to hold their foliage well. Best known is the large-leaved *F. elastica* 'Decora,' but perhaps even more attractive, because of their very graceful habit, are several small-leaved kinds, such as *F. benjamina, F. retusa,* and *F. nitida.* The giant violin-like, leathery leaves of *F. lyrata,* better known as *F. pandurata,* make the plant an attractive indoor "tree." *Coccoloba,* the sea grape, is another sturdy woody plant, somewhat resembling *Ficus,* with leathery, rounded leaves and crimson veining.

Because of their majestic beauty and distinctive decorative appeal many palms are grown indoors. Best known of the feather palms is the paradise palm (*Howea,* or *Kentia*), which combines grace with sturdiness; its thick, leathery leaves can stand much abuse. The parlour palms and bamboo palms of the genus *Chamaedorea* have dainty fronds on slender stalks; they keep well even in fairly dark places. Similar in appearance is the areca palm (*Chrysalidocarpus*) with slender yellowish stems carrying feathery fronds in clusters. The pygmy date (*Phoenix roebelenii*), a compact palm with gracefully arching, dark-green leaves, is an excellent houseplant if kept warm and moist.

Ferns. Ferns, which come in a wide variety of forms, provide many popular houseplants. Among the best smaller parlour ferns is the sword fern, *Nephrolepis,* with bushy rosettes of leafy fronds; the holly fern (*Cyrtomium*), which has glossy dark leathery leaves; and the leather-leaf fern (*Rumohra*), with its leathery but lacy fronds. The bird's-nest fern (*Asplenium nidus*) forms a rosette of parchment-textured, fanlike, light-green fronds. Long-lasting *Polypodium,* often known as hare's-foot fern because of its pawlike, woolly rhizomes (rootlike structures), has feathery leaves on slender stalks. Among the attractive damp-loving ferns are the several species of dainty maidenhairs (*Adiantum*). The so-called table ferns are a varied group of mainly *Pteris* and *Pellaea* species; some are frilly, others variegated; and in their younger stages they are ideal subjects for terraria. The *Platycerium,* or staghorn fern, has always aroused great curiosity because of its unusual shape. Growing as epiphytes on trees, these ferns have sterile fronds that cling snugly to the bark or, in cultivation, to a wire basket or wooden block; their much divided fertile fronds resemble the antlers of deer. One of the best of the palmlike tree ferns is the Hawaiian *Cibotium,* with a stout, fibrous trunk that bears a crown of light-green fronds.

Popular fernlike plants include *Asparagus* species that have plumy fronds. Species of *Selaginella,* called sweat plants or club moss, are strictly warm terrarium subjects; their delicate fronds greedily soak up moisture from the atmosphere to keep from shriveling.

Climbers and trailers. Climbers and trailers, weeping plants with stems too weak to support themselves, occur in most plant families. Best known are many varieties of ivy (*Hedera*). Generally, they prefer a cool location, but some small-leaved or variegated varieties do well on the windowsill. Several *Cissus* species, such as *C. rhombifolia,* the grape ivy, with metallic foliage, and the leathery *C. antarctica,* or kangaroo vine, are excellent plants for boxes or room dividers. Intriguing is the slow-growing *Hoya,* or wax plant, with leathery foliage and waxy, wheel-shaped blooms. By contrast, the inch plants and wandering jew, species of *Tradescantia* and *Zebrina,* are rapid growers with watery stems and varicoloured leaves; these long-beloved houseplants are used widely in window shelves or hanging baskets. The spider plants (*Chlorophytum,* or *Anthericum*) are houseplant favourites, forming clusters of fresh green ribbonlike leaves banded white; young plantlets develop from the tips of arching stalks.

Flowering plants. Most of the flowering potted plants seen at holiday times are not easy subjects for long-term indoor cultivation. They require high light intensity, careful watering, and day–night differences in temperature that are not usually available in the home; greenhouses offer better chances for successful cultivation. There are exceptions, however; one of the most successfully adapted houseplants is the African violet (*Saintpaulia*), with countless named varieties, with blossoms from violet blue through rose to white and single- and double-flowered forms. Window bloomers, such as *Abutilon,* the parlour maples, have bell-like flowers resembling Chinese lanterns. *Impatiens,* or busy Lizzie, is a genus of succulent herbs producing a succession of spurred flowers in gay colours. *Hibiscus,* the rose mallows, has short-lived giant blossoms in brilliant colours. Geraniums (botanically *Pelargonium*) have long been popular flowering plants in the sunny window; the foliage is often variegated or scented, and flower clusters may be in reds, pinks, and white.

A number of bulbous plants do well in lighted windows: *Hippeastrum,* better known as amaryllis; *Clivia,* the Kaffir lily; *Haemanthus,* the blood lily; *Neomarica,* the apostle plant; and *Veltheimia,* the forest lily.

Orchids present a more difficult and specialized subject for successful home cultivation, usually because of their requirements for light, controlled temperature, and sufficient humidity and ventilation. There are some kinds, however, that give good results with ordinary care: epiphytic *Epidendrum* species, with waxy, usually fragrant, often greenish blossoms; and *Oncidium* species, or butterfly orchids, with brightly coloured, long-lasting yellow flowers marked with brown and often produced in large sprays.

Small fruit plants

Small flowering plants that produce edible fruit can be grown on a windowsill. With sufficient light and ventilation, success may be had with the Calamondin orange (×*Citrofortunella mitis*), the dwarf Chinese lemon (*Citrus limon* 'Meyeri'), and the American-wonder lemon (*C. limon* 'Ponderosa'). The fig tree (*Ficus carica*) can be grown to yield edible fruit, as can the dwarf Cavendish banana (*Musa acuminata,* formerly *M. nana*) and the dwarf pomegranate (*Punica granatum nana*), the pineapple (*Ananas comosus*), and the coffee tree (*Coffea arabica*).

(A.B.Gr./Ro.P.)

Horticulture

Horticulture is the branch of plant agriculture dealing with garden crops, generally fruits, vegetables, and ornamental plants. The word is derived from the Latin *hortus,* "garden," and *colere,* "to cultivate." As a general term, it covers all forms of garden management; but in ordinary use it refers to intensive commercial production. In terms of scale, horticulture falls between domestic gardening and field agriculture, though all forms of cultivation naturally have close links.

Horticulture is divided into the cultivation of plants for food (pomology and olericulture) and plants for ornament (floriculture and landscape horticulture). Pomology deals with fruit and nut crops. Olericulture deals with herbaceous plants for the kitchen, including, for example, carrots (edible root), asparagus (edible stem), lettuce (edible leaf), cauliflower (edible flower), tomatoes (edible fruit), and peas (edible seed). Floriculture deals with the production of flowers and ornamental plants; generally, cut flowers, pot plants, and greenery. Landscape horticulture is a broad category that includes plants for the landscape, including lawn turf, but particularly nursery crops such as shrubs, trees, and climbers.

The specialization of the horticulturist and the success of the crop are influenced by many factors. Among these are climate, terrain, and other regional variations. (Ro.P.)

HORTICULTURAL REGIONS

Temperate zones. Temperate zones for horticulture cannot be defined exactly by lines of latitude or longitude but are usually regarded as including those areas where frost in winter occurs, even though rarely. Thus most parts of Europe, North America, and northern Asia are included, though some parts of the United States, such as southern California and Florida, are considered subtropical. A few parts of the north coast of the Mediterranean and the Mediterranean islands are also subtropical. In the Southern Hemisphere, practically all of New Zealand, a few parts of Australia, and the southern part of South America have temperate climates. For horticultural purposes altitude is also a factor; the lower slopes of great mountain ranges, such as the Himalayas and the Andes, are included. Thus the temperate zones are very wide and the range of plants that can be grown in them is enormous, probably greater than in either the subtropical or tropical zones. In the temperate zones are the great coniferous and deciduous forests: pine, spruce, fir, most of the cypresses, the deciduous oaks (but excluding many of the evergreen ones), ash, birch, and linden (lime).

The temperate zones are also the areas of the grasses—the finest lawns particularly are in the regions of moderate or high rainfall—and of the great cereal crops. Rice is excluded as being tropical, but wheat, barley, corn (maize), and rye grow well in the temperate zones.

The winter rest

Plants in the temperate zones benefit from a winter resting season, which clearly differentiates them from tropical plants, which tend to grow continuously. Bulbs, annuals, herbaceous perennials, and deciduous trees become more frost-resistant with the fall of sap and therefore have a better chance of passing the resting season undamaged. Another influence is the varying length of darkness and light throughout the year, so that many plants, such as chrysanthemums, have a strong photoperiodism. The chrysanthemum flowers only in short daylight periods, although artificial lighting in nurseries can produce flowers the year round.

Most of the great gardens of the world have been developed in temperate zones. Particular features such as rose gardens, herbaceous borders, annual borders, woodland gardens, and rock gardens are also those of temperate-zone gardens. Nearly all depend for their success on the winter resting period.

Tropical zones. There is no sharp line of demarcation between the tropics and the subtropics. Just as many tropical plants can be cultivated in the subtropics, so also many subtropical and even temperate plants can be grown satisfactorily in the tropics. Elevation is a determining factor. For example, the scarlet runner bean, a common plant in temperate regions, grows, flowers, and develops pods normally on the high slopes of Mt. Meru in Africa near the Equator; it will not, however, set pods in Hong Kong, a subtropical situation a little south of the Tropic of Cancer but at a low elevation.

Effect of rainfall and length of daylight

In addition to elevation, another determinant is the annual distribution of rainfall. Plants that grow and flower in the monsoon areas, as in India, will not succeed where the climate is uniformly wet, as in Bougainville in the Solomon Islands. Another factor is the length of day, the number of hours the Sun is above the horizon; some plants flower only if the day is long, but others make their growth during the long days and flower when the day is short. Certain strains of the cosmos plant are so sensitive to light that, where the day is always about 12 hours, as near the Equator, they flower when only a few inches high; if grown near the Tropics of Cancer or Capricorn, they attain a height of several feet, if the seeds are sown in the spring, before flowering in the short days of autumn and winter. Poinsettia is a short-day plant that may be seen in flower in Singapore on any day of the year, while in Trinidad it is a blaze of glory only in late December.

In the tropics of Asia and parts of Central and South America the dominant features of the gardens are flowering trees, shrubs, and climbers. Herbaceous plants are relatively few, but many kinds of orchids can be grown.

Vegetable crops vary in kind and quality with the presence or absence of periodic dry seasons. In the uniformly wet tropics, the choice is limited to a few root crops and still fewer greens. Sweet potatoes grow and bear good crops where the average monthly rainfall, throughout the year, exceeds 10 inches (25 centimetres); they grow even better where there is a dry season. The same can be said of taro, yams, and cassava. Tropical greens from the Malay Peninsula are not as good as those grown in South China, the Hawaiian Islands, and Puerto Rico. They include several spinaches, of which Chinese spinach or amaranth is the best; several cabbages; Chinese onions and chives; and several gourds, cucumbers, and, where there is a dry season, watermelons. Brinjals, or eggplants, peppers, and okra are widely cultivated. Many kinds of beans can be grown successfully, including the French bean from the American subtropics, the many varieties of the African cowpea, and yard-long bean. The yam bean, a native of tropical America, is grown for its edible tuber. In the drier areas the pigeon pea, the soybean, the peanut (groundnut), and the Tientsin green bean are important crops. Miscellaneous crops include watercress, ginger, lotus, and bamboo. (G.A.C.H./P.M.Sy./Ed.)

PROPAGATION

Propagation, the controlled perpetuation of plants, is the most basic of horticultural practices. Its two objectives are to achieve an increase in numbers and to preserve the essential characteristics of the plant. Propagation can be achieved sexually by seed or asexually by utilizing specialized vegetative structures of the plant (tubers and corms) or by employing such techniques as cutting, layering, grafting, and tissue culture. (A detailed discussion of the methods of controlling sexual propagation can be found in the article FARMING AND AGRICULTURAL TECHNOLOGY.)

Seed propagation. The most common method of propagation for self-pollinated plants is by seed. In self-pollinated plants, the sperm nuclei in pollen produced by a flower fertilize egg cells of a flower on the same plant. Propagation by seed is also used widely for many cross-pollinated plants (those whose pollen is carried from one plant to another). Seed is usually the least expensive and often the only means of propagation and offers a convenient way to store plants over long periods of time. Seed kept dry and cool normally maintains its viability from harvest to the next planting season. Some can be stored for years under suitable conditions. Seed propagation also makes it possible to start plants free of most diseases. This is especially true with respect to virus diseases, because it is almost impossible to free plants of virus infections and because most virus diseases are not transmitted by seed. There are two disadvantages to seed propagation. First, genetic variation occurs in seed from cross-pollinated plants because they are heterozygous. This means that the plant grown from seed may not exactly duplicate the characteristics of its parents and may possess undesirable characteristics. Second, some plants take a long time to grow from seed to maturity. Potatoes, for example, do not breed true from seed and do not produce large tubers the first year. These disadvantages are overcome by vegetative propagation.

The practice of saving seed to plant the following year has developed into a specialized part of horticulture. Seed technology involves all of the steps necessary to ensure production of seed with high viability, freedom from disease, purity, and trueness to type. These processes may include specialized growing and harvesting techniques, cleaning, and distribution.

Relatively little tree and shrub seed is grown commercially; it is generally harvested from natural stands. Rootstock seed for fruit trees is often obtained as a by-product in fruit-processing industries. Seed growing and plant improvement are related activities. Thus many seed-producing firms actively engage in plant-breeding programs to accomplish genetic improvement of their material.

Harvesting of dry seed is accomplished by threshing. Seed from fleshy fruits is recovered through fermentation of the macerated (softened by soaking) pulp or directly from screening. Machines have been developed to separate and clean seed, based on size, specific gravity, and surface characteristics. Extended storage of seed requires low humidity and cool temperature.

Trade in seed requires quality control. For example, U.S. government seed laws require detailed labeling showing germination percentage, mechanical purity, amount of seed, origin, and moisture content. Seed testing is thus an important part of the seed industry.

While most vegetable seed germinates readily upon exposure to normally favourable environmental conditions, many seed plants that are vegetatively (asexually) propagated fail to germinate readily because of physical or physiologically imposed dormancy. Physical dormancy is due to structural limitations to germination such as hard impervious seed coats. Under natural conditions weathering for a number of years weakens the seed coat. Certain seeds, such as the sweet pea, have a tough husk that can be artificially worn or weakened to render the seed coat permeable to gases and water by a process known as scarification. This is accomplished by a number of methods including abrasive action, soaking in hot water, or acid treatment. Physiologically imposed dormancy involves the presence of germination inhibitors. Germination in such seed may be accomplished by treatment to remove these inhibitors. This may involve cold stratification, storing seed at high relative humidity and low temperatures, usually slightly above freezing. Cold stratification is a prerequisite to the uniform germination of many temperate-zone species such as apple, pear, and redbud.

Scarification

Vegetative propagation. Asexual or vegetative reproduction is based on the ability of plants to regenerate tissues and parts. In many plants vegetative propagation is a completely natural process; in others it is an artificial one. Vegetative propagation has many advantages. These include the unchanged perpetuation of naturally cross-pollinated or heterozygous plants and the possibility of propagating seedless progeny. This means that a superior plant may be reproduced endlessly without variation. In addition, vegetative propagation may be easier and faster than seed propagation, because seed dormancy problems are eliminated and the juvenile nonflowering stage of some seed-propagated plants is eliminated or reduced.

Vegetative propagation is accomplished by use of (1) apomictic seed; (2) specialized vegetative structures such as runners, bulbs, corms, rhizomes, offshoots, tubers, stems, and roots; (3) layers and cuttings; (4) grafting and budding; and (5) tissue culture.

Apomixis. Apomixis, the development of asexual seed (seed not formed via the normal sexual process), is a form of vegetative propagation for some horticultural plants including Kentucky bluegrass, mango, and citrus. Virus-free progeny can be produced in oranges from a seed that is formed from the nucellus, a maternal tissue.

Vegetative structures. Many plants produce specialized vegetative structures that can be used in propagation. These may be storage organs such as tubers that enable the plant to survive adverse conditions or organs adapted for natural propagation—runners or rhizomes—so that the plant may rapidly spread.

Bulbs and corms

Bulbs consist of a short stem base with one or more buds protected by fleshy leaves. They are found in such plants as the onion, daffodil, and hyacinth. Bulbs commonly grow at ground level, though bulblike structures (bulbils) may form on aerial stems in some lilies or in association with flower parts, as in the onion. Buds in the axils (angle between leaf and stem) of the fleshy leaves may form miniature bulbs (bulblets) that when grown to full size are known as offsets. Corms are short, fleshy, underground stems without fleshy leaves. The gladiolus and crocus are propagated by corms. They may produce new cormels from fleshy buds. Rhizomes are horizontal, underground stems that are compressed, as in the iris, or slender, as in turf grasses. Runners are specialized aerial stems, a natural agent of increase and spread for such plants as the strawberry, strawberry geranium, and bugleweed (*Ajuga*). Tubers are fleshy enlarged portions of underground stem. The edible portion of the potato, the tuber, is also used as a means of propagation.

A number of plants form lateral shoots from the stem, which when rooted serve to propagate the plant. These are known collectively as offshoots but are often called offsets, crown divisions, ratoons, or slips.

Roots may also be structurally modified as propagative and food-storage organs. These tuberous roots, fleshy swollen structures, readily form shoots (called adventitious, because they do not form from nodes). The sweet potato and dahlia are propagated by tuberous roots. Shoots that rise adventitiously from roots are called suckers. The red raspberry is propagated by suckers.

Layering and cutting. Propagation can be accomplished by methods in which plants are induced to regenerate missing parts, usually adventitious roots or shoots. When the regenerated part is still attached to the plant the process is called layerage, or layering; when the regenerating portion is detached from the plant the process is called cuttage, or cutting.

Layering often occurs naturally. Drooping black raspberry stems tend to root in contact with the soil. The croton, a tropical plant, is commonly propagated by wrapping moist sphagnum enclosed in plastic around a stem cut to induce rooting. After rooting, the stem is detached and planted. Though simple and effective, layering is not normally adapted to large-scale nursery practices.

Cutting is one of the most important methods of propagation. Many plant parts can be used; thus cuttings are classified as root, stem, or leaf. Stem cuttings are the most common.

The ability of stems to regenerate missing parts is variable; consequently plants may be easy or difficult to root. The physiological ability of cuttings to form roots is due to an interaction of many factors. These include transportable substances in the plant itself: plant hormones (such as auxin), carbohydrates, nitrogenous substances, vitamins, and substances not yet identified. Environmental factors such as light, temperature, humidity, and oxygen are important, as are age, position, and type of stem.

Although easy-to-root plants such as willow or coleus can be propagated merely by plunging a stem in water or moist sand, the propagation of difficult-to-root species is a highly technical process. To achieve success with difficult-to-root plants special care is taken to control the environment and encourage rooting. A number of growth regulators stimulate rooting. A high degree of success has been achieved with indolebutyric acid, a synthetic auxin that is applied to the cut surface. A number of materials known as rooting cofactors have been found that interact with auxin to further stimulate rooting, and these are sold as a hormone rooting compound.

Humidity control is particularly important to prevent death of the stem from desiccation before rooting is complete. The use of an intermittent-mist system in propagation beds has proved to be an important means of improving success in propagation by cuttings. These operate by applying water to the plant for a few seconds each minute.

Grafting. Grafting involves the joining together of plant parts by means of tissue regeneration. The part of the combination that provides the root is called the stock; the added piece is called the scion. When more than two parts are involved, the middle piece is called the interstock. When the scion consists of a single bud, the process is called budding. Grafting and budding are the most widely used of the vegetative propagation methods.

How grafting works

Stock cambium and scion cambium respond to being cut by forming masses of cells (callus tissues) that grow over the injured surfaces of the wounds. The union resulting from interlocking of the callus tissues is the basis of graftage. In dicots (*e.g.,* most trees) cambium—a layer of actively dividing cells between xylem (wood) and phloem (bast) tissues—is usually arranged in a continuous ring; in woody members, new layers of tissue are produced annually. Monocot stems (*e.g.,* lilacs, orchids) do not possess a continuous cambium layer or increase in thickness; grafting is seldom possible.

The basic technique in grafting consists of placing cambial tissues of stock and scion in intimate association, so that the resulting callus tissue produced from stock and scion interlocks to form a living continuous connection. A snug fit can be obtained through the tension of the split stock and scion or both. Tape, rubber, and nails can be used to achieve close contact. In general, grafts are only compatible between the same or closely related species. Success in grafting depends on skill in achieving a snug fit. Warm temperatures (80°–85° F [27°–30° C]) increase callus formation and improve "take" in grafting. Thus grafts using dormant material are often stored in a warm, moist place to stimulate callus formation.

In grafting and budding, the rootstock can be grown

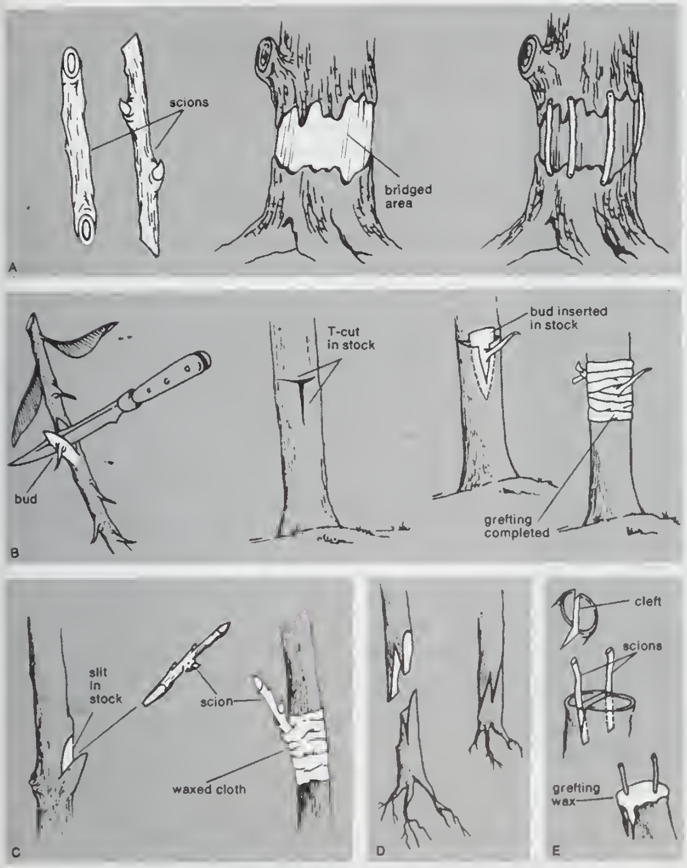


Figure 2: *Methods of grafting and budding.*
(A) Bridge graft of a damaged tree. (B) Budding. (C) Side graft. (D) Whip-and-tongue graft. (E) Cleft graft.

From *Plant Science: An Introduction to World Crops* by Jules Janick, Robert W. Schery, Frank W. Woods, and Vernon W. Ruttan. W.H. Freeman and Company. Copyright © 1969

from seed or propagated asexually. Within a year a small amount of scion material from one plant can produce hundreds of plants. Some methods of grafting and budding are illustrated in Figure 2.

Grafting has uses in addition to propagation. The interaction of rootstocks may affect the performance of the stock through dwarfing or invigoration and in some cases may affect quality. Further, the use of more than one component can affect the disease resistance and hardiness of the combination.

John H. Gerard

Healed cleft graft on cherry (*Prunus*) stock. The scions (top) are double flowering cherry.

Grafting as a means of growth control is used extensively with fruit trees and ornamentals such as roses and junipers. Fruit trees are normally composed of a scion grafted onto a rootstock. Sometimes an interstock is included between the scion and stock. The rootstock may be grown from seed (seedling rootstock) or asexually propagated (clonal rootstock). In the apple, a great many clonal rootstocks are available to give a complete range of dwarfing; rootstocks are also available to invigorate growth of the scion cultivar.

Tissue-culture techniques utilizing embryos, shoot tips, and callus can be used as a method of propagation. The procedure requires aseptic techniques and special media to supply inorganic elements; sugar; vitamins; and, depending on the tissue, growth regulators and organic complexes such as coconut milk, yeast, and amino-acid extract.

Embryo culture has been used to produce plants from embryos that would not normally develop within the fruit. This occurs in early-ripening peaches and in some hybridization between species. Embryo culture can also be used to circumvent seed dormancy.

A shoot tip, when excised and cultured, may produce roots at the base. This technique is employed for the purpose of producing plants free of disease. Certain orchids are rapidly multiplied by this method. Cultured shoot tips form an embryo-like stage that can be sectioned indefinitely to build up large stocks rapidly. These bulblike bodies left unsectioned develop into small plantlets. A similar procedure is used with the carnation, in which the shoot tip forms a cell mass that can be subdivided.

Callus-tissue culture—a very specialized technique that involves growth of the callus, followed by procedures to induce organ differentiation—has been successful with a number of plants including carrot, asparagus, and tobacco. Used extensively in research, callus culture has not been considered a practical method of propagation. Callus culture produces genetic variability because in some cases cells double their chromosome number. In rice and tobacco, mature plants have been obtained from callus formed from pollen. These plants have half the normal number of chromosomes.

BREEDING

The isolation and production of superior types known as cultivars are the very keystones of horticulture. Plant breeding, the systematic improvement of plants through the application of genetic principles, has placed improvement of horticultural plants on a scientific basis. The raw material of improvement is found in the great variation that exists between cultivated plants and related wild species. The incorporation of these changes into cultivars adapted to specific geographical areas requires a knowledge of the theoretical basis of heredity and art and the skill to discover, perpetuate, and combine these small but fundamental differences in plant material.

The goal of the plant breeder is to create superior crop varieties. The cultivated variety, or cultivar, can be defined as a group of crop plants having similar but distinguishable characteristics. The term cultivar has various meanings, however, depending on the mode of reproduction of the crop. With reference to asexually propagated crops, the term cultivar means any particular clone considered of sufficient value to be graced with a name. With reference to sexually propagated crops, the concept of cultivar depends on the method of pollination. The cultivar in self-pollinated crops is basically a particular homozygous genotype, a pure line. In cross-pollinated crops the cultivar is not necessarily typified by any one plant but sometimes by a particular plant population, which at any one time is composed of genetically distinguishable individuals.

ENVIRONMENTAL CONTROL

Control of the natural environment is a major part of all forms of cultivation, whatever its scale. The basic processes involved in this task have already been described in a preceding section on the principles of gardening, and these also apply to horticulture. The scale, intensiveness, and economic risk in commercial gardening and nurs-

Poinsettias (*Euphorbia pulcherrima*) being cultivated in the controlled environment of a modern greenhouse.

John H Gerard

eries, however, often require approaches markedly different from those of the small home garden; and some of these are described here.

The intensive cultivation practiced in horticulture relies on extensive control of the environment for all phases of plant life. The most basic environmental control is achieved by location and site: sunny or shady sites, proximity to bodies of water, altitude, and latitude.

Structures. Various structures are used for temperature control. Cold frames, used to start plants before the normal growing season, are low enclosed beds covered with a removable sash of glass or plastic. Radiant energy passes through the transparent top and warms the soil directly. Heat, however, as long-wave radiation, is prevented from leaving the glass or plastic cover at night. Thus heat that builds up in the cold frame during the day aids in warming the soil, which releases its heat gradually at night to warm the plants. When supplemental heat is provided, the structures are called hotbeds. At first, supplemental heat was supplied by respiration through the decomposition of manure or other organic matter. Today, heat is provided by electric cables, steam, or hot-water pipes buried in the soil.

Greenhouses

Greenhouses are large hotbeds, and in most cases the source of heat is steam. While they were formerly made of glass, plastic films are now extensively used. Modern greenhouse ranges usually have automatic temperature control. Summer temperatures can be regulated by shading or evaporative "fan-and-pad" cooling devices. Air-conditioning units are usually too expensive except for scientific work. Greenhouses with precise environmental controls are known as phytotrons. Other environmental factors are controlled through automatic watering, regulation of light and shade, addition of carbon dioxide, and the regulation of fertility.

Shade houses are usually walk-in structures with shading provided by lath or screening. Summer propagation is often located in shade houses to reduce excessive water loss by transpiration.

Temperature control. A number of temperature-control techniques are used in the field, including application of hot caps, cloches, plastic tunnels, and mulches of various types. Hot caps are cones of translucent paper or plastic that are placed over the tops of plants in the spring. These act as miniature greenhouses. In the past small glass sash called cloches were placed over rows to help keep them warm. Polyethylene tunnels supported by wire hoops that span the plants are now used for the same purpose. As spring advances the tunnels are slashed to prevent excessive heat buildup. In some cases the plastic tunnels are constructed so that they can be opened and closed when necessary. This technique is widely used in Israel for early production of vegetables.

Mulching (already described in its application to domestic gardens) is important in horticulture. Whether in the form of a topdressing of manure or compost or plastic sheeting, mulches offer the grower the various benefits of economical plant feeding, conservation of moisture, and control of weeds and erosion. Winter mulches are commonly used to protect such sensitive and valuable plants as strawberries and roses.

The storage of perishable plant products is accomplished largely through the regulation of their temperature to retard respiration and microbial activity. Excess water loss can be prevented by controlling humidity. Facilities that utilize the temperature of the atmosphere are called common storage. The most primitive types take advantage of the reduced temperature fluctuations of the soil by using caves or unheated cellars. Aboveground structures must be insulated and ventilated. Complete temperature-regulated storages utilizing refrigeration and heating are now common for storage of horticultural products. The regulation of oxygen and carbon dioxide levels along with the regulation of temperature is known as controlled-atmosphere storage. Rooms are sealed so that gaseous exchange can be effectively controlled. Many horticultural products, such as fruit, can be kept fresh for as long as a year under these controlled conditions.

Frost control. Frost is one of the high-risk elements for commercial growers, and the problem is accentuated by the fact that growers are striving to produce early-season crops. The precautions are consequently far more elaborate and costly than those of the domestic garden. Frost is especially damaging to perennial fruit crops in the spring—because flower parts are sensitive to freezing injury—and to tender transplants. The two weather conditions that produce freezing temperatures are rapid radiational cooling at night and introduction of a cold air mass with temperatures below freezing. Radiation frost occurs when the weather is clear and calm; air-mass freezes occur when it is overcast and windy.

Frost-control methods involve either reduction of radiational heat loss or conservation or addition of heat. Radiational heat loss may be reduced by hot caps, cold frames, or mulches. Heat may also be added from the air. Wind machines that stir up the air, for example, provide heat when temperature inversions trap cold air under a layer of warm air. These have been used extensively in citrus groves. Heat may be added directly by using heaters, usually fueled with oil. Sprinkler irrigation can also be used for frost control. The formation of ice is accompanied by the release of large amounts of heat, which maintains plants at the freezing temperature as long as the water is being frozen. Thus continuous sprinkling during frosty nights has been used to protect strawberries from frost injury.

Frost injury to transplants can be prevented through processes that increase the plant's ability to survive the impact of unfavourable environmental stress. This is known as hardening off. Hardening off of plants prior to transplanting can be accomplished by withholding water and fertilizer, especially nitrogen. This prevents formation of succulent tissue that is very frost-tender. Gradual exposure to cold is also effective for hardening. Induced cold resistance in crops such as cabbage, for example, can have a considerable effect; unhardened cabbages begin to show injury at 28° F (−2.2° C), while hardened plants withstand temperatures as low as 22° F (−5.6° C).

Light control. Light has a tremendous effect on plant growth. It provides energy for photosynthesis, the process by which plants, with the aid of the pigment chlorophyll, synthesize carbon compounds from water and carbon dioxide. Light also influences a great number of physiological reactions in plants. At energy values lower than those required for photosynthesis, light affects such processes as dormancy, flowering, tuberization, and seed-stalk development. In many cases these processes are affected by the length of day; the recurrent cycle of light is known as the photoperiod.

The control of light in horticultural practices involves increasing energy values for photosynthesis and controlling day length. Light is controlled in part by site and location. In the tropics day length approaches 12 hours throughout the year, whereas in polar regions it varies from zero to 24 hours. Light is also partly controlled by plant distribution and density.

Supplemental illumination in greenhouses increases photosynthesis. The cost of power to supply the artificial light, however, makes this impractical for all but crops of the highest value. Fluorescent lights are the most efficient for photosynthesis; special lights, rich in the wavelengths required, are now available.

Extension of day length through supplemental illumination and shading is common practice in the production of greenhouse flower crops, which are often induced to flower out of season. Artificial lengthening of short days, or interruption of the dark period, promotes flowering in long-day plants such as lettuce and spinach and prevents flowering of short-day plants such as chrysanthemums. Similarly, during naturally long days, shading to reduce day length prevents flowering of long-day plants and promotes flowering of short-day plants. The manipulation of day length is standard practice to control flowering of greenhouse chrysanthemums throughout the year. Tungsten lights have proved very effective for extending day length because they are rich in the red end of the spectrum that affects the photoperiodic reaction. Extending the day length is a relatively affordable practice because only a low light intensity is required. The same effects can be obtained through interruption of the dark period, even with light flashes. Decreasing day length is usually accomplished by simply covering the plants with black shade cloth.

Soil management. The principles involved here are again similar to those of home gardening. But the financial considerations of horticulture naturally require a more scientific approach to soil care. To be successful, the grower must ensure the economic use of every square yard of ground, especially because the cost of sound horticultural land is among the highest of any in agriculture. Crop rotation is planned to ensure that the soil is not depleted of essential chemicals by repeated use of one type of plant in the same plot. Soil analysis is employed so that any such depletion can be rectified promptly. Fertilizers are applied in a precise routine and, of course, in a variety beyond the reach or needs of the ordinary gardener. They are frequently applied through leaves or stems in the form of chemical sprays.

Water management. Depending on the terrain, water management may involve extensive works for irrigation and drainage. While the home gardener may well be content with a rough-and-ready appraisal of the wetness or dryness of the soil, horticulture is more exacting. Production of the high-quality fruits and vegetables demanded by the modern market requires a precise all-year balance of soil moisture, adjusted to the needs of the particular crop. These considerations apply whether the grower is situated in a high-rainfall area of Europe or in the parched land of the southwestern United States or Israel.

There are a number of general methods of land irrigation. In surface irrigation water is distributed over the surface of soil. Sprinkler irrigation is application of water under pressure as simulated rain. Subirrigation is the distribution of water to soil below the surface; it provides moisture to crops by upward capillary action. Trickle irrigation involves the slow release of water to each plant through small plastic tubes. This technique is adapted both to field and to greenhouse conditions.

Removal of excess water from soils can be achieved by surface or subsurface drainage. Surface drainage refers to the removal of surface water by development of the slope of the land utilizing systems of drains to carry away the surplus water. In subsurface drainage open ditches and tile fields intercept groundwater and carry it off. The water enters the tiling through the joints, and drainage is achieved by gravity feed through the tiles.

Pest control. Horticultural plants are subject to a wide variety of injuries caused by other organisms. Plant pests include viruses, bacteria, fungi, higher plants, nematodes, insects, mites, birds, and rodents. Various methods are used to control them. The most successful treatments are preventive rather than curative.

Control of pests is achieved through practices that prevent harm to the plant and methods that affect the plant's ability to resist or tolerate intrusion by the pathogen. These can be classified as cultural, physical, chemical, or biological.

Traditional practices that reduce effective pest population include the elimination of diseased or infected plants or seeds (roguing), cutting out of infected plant parts (surgery), removal of plant debris that may harbour pests (sanitation), and alternating crops unacceptable to pests (rotation). Any of a number of techniques can be employed to render the environment unfavourable to the pest, such as draining or flooding and changing the soil's level of acidity or alkalinity.

Physical methods can be used to protect the plant against intrusion or to eliminate the pest entirely. Physical barriers range from the traditional garden fence to bags that protect each fruit, a common practice in Japan. Heat treatment is used to destroy some seed-borne pathogens and is a standard soil treatment in greenhouses to eliminate soil pests such as fungi, nematodes, and weed seed. Cultivation and tillage are standard practices for weed control.

The horticultural industry is now dependent upon chemical control of pests through pesticides, materials toxic to the pest in some stage of its life cycle. Commercial growers of practically all horticultural crops rely on complete schedules utilizing many different compounds. Pesticides are usually classed according to the organism they control: for example, bactericide, fungicide, nematicide, miticide, insecticide, rodenticide, and herbicide.

Pesticide selectivity

Selectivity of pesticides, the ability to discriminate between pests, is a relative concept. Some nonselective pesticides kill indiscriminately; most are selective to some degree. Most fungicides, for example, are not bactericidal. The development of highly selective herbicides makes it possible to destroy weeds from crops selectively. Selectivity can be achieved through control of dosage, timing, and method of application.

Plant pests can also be controlled through the manipulation of biological factors. This may be achieved through directing the natural competition between organisms or by incorporating natural resistance to the whole plant. The introduction of natural parasites or predators has been a successful method for the control of certain insects and weeds. Incorporation of genetic resistance is an ideal method of control. Thus breeding for disease and insect resistance is one of the chief goals of plant breeding programs. A major obstacle to this method of control is the ability of pathogens (disease-producing organisms) to mutate easily and attack previously resistant plants.

GROWTH REGULATION BY CHEMICALS

Control of plant growth through growth-regulating materials is a modern development in horticulture. These materials have resulted from basic investigations into growth and development, as well as systematic screening of materials to find those that affect differentiation and growth. This field was given great impetus by the discovery of a class of plant hormones known as auxins, which affect cell elongation.

Auxins have been correlated with inhibition and stimulation of growth as well as differentiation of organs and tissues. Such processes as cell enlargement, leaf and organ separation, budding, flowering, and fruit set (the formation of the fruit after pollination) and growth are influenced by auxins. In addition, auxins have been associated with the movement of plants in response to light and gravity. Auxin materials are used in horticulture for the promotion of rooting, fruit setting, fruit thinning, and fruit-drop control.

Gibberellins are a group of related, naturally occurring compounds of which only one, gibberellic acid, is commercially available. Gibberellins have many effects on plant development. The most startling is the stimulation of growth in many compact or dwarf plants. Minute applications transform bush to pole beans or dwarf to normal corn. Perhaps the most widespread horticultural use has been in grape production. The application of gibberellin is now a regular practice for the culture of the 'Thompson seedless' cultivar ("Sultanina") of grapes to increase berry size. In Japan applications of gibberellic acid are used to induce seedlessness in certain grapes.

Cytokinins are a group of chemical substances that have a decisive influence on the stimulation of cell division. In tissue culture high auxin and low cytokinin give rise to root development; low auxin and high cytokinin encourage shoot development.

Ethylene, a hydrocarbon compound, acts as a plant hormone to stimulate fruit ripening as well as rooting and flowering of some plants. An ethylene-releasing compound, 2-chloroethylphosphonic acid, has many horticultural applications, of which the most promising may be uniform ripening of tomatoes and the stimulation of latex flow in rubber.

Many compounds that inhibit growth hormones have application in horticulture. For example, a number of materials that inhibit formation of gibberellins by the plant cause dwarfing. These include chlorinated derivatives of quaternary ammonium and phosphonium compounds. Many of these have applications in floriculture. Growth retardants such as succinic acid–2,2-dimethylhydrazide, a gibberellin suppressor, have applications in horticulture from a wide array of effects that include dwarfing and fruit maturity. The growth inhibitor maleic hydrazide has been effective in preventing the sprouting of onions and potatoes.

ORNAMENTAL HORTICULTURE

Ornamental horticulture consists of floriculture and landscape horticulture. Each is concerned with growing and marketing plants and with the associated activities of flower arrangement and landscape design. The turf industry is also considered a part of ornamental horticulture. Although flowering bulbs and flower seed represent an important component of agricultural production for the Low Countries of Europe, ornamentals are relatively insignificant in world trade.

Floriculture has long been an important part of horticulture, especially in Europe and Japan, and accounts for about half of the nonfood horticultural industry in the United States. Because flowers and pot plants are largely produced in plant-growing structures in temperate climates, floriculture is largely thought of as a greenhouse industry; there is, however, considerable outdoor culture of many flowers.

The industry is usually very specialized with respect to its crop; the grower must provide precise environmental control. Exact scheduling is imperative since most floral crops are seasonal in demand. Because the product is perishable, transportation to market must function smoothly to avoid losses.

The floriculture industry involves the grower, who mass-produces flowers for the wholesale market, and the retail florist, who markets to the public. The grower is often a family farm, but, as in all modern agriculture, the size of the growing unit is increasing. There is a movement away from urban areas, with their high taxes and labour costs, to locations with lower tax rates and a rural labour pool and also toward more favourable climatic regions (milder temperature and more sunlight). The development of airfreight has emphasized interregional and international competition. Flowers can be shipped long distances by air and arrive in fresh condition to compete with locally grown products.

The industry of landscape horticulture is divided into growing, maintenance, and design. Growing of plants for landscape is called the nursery business, although a nursery refers broadly to the growing and establishment of any young plant before permanent planting. The nursery industry involves production and distribution of woody and herbaceous plants and is often expanded to include ornamental bulb crops—corms, tubers, rhizomes, and swollen roots as well as true bulbs. Production of cuttings to be grown in greenhouses or for indoor use (foliage plants), as well as the production of bedding plants, is usually considered part of floriculture, but this distinction is fading. While most nursery crops are ornamental, the nursery business also includes fruit plants and certain perennial vegetables used in home gardens, for example, asparagus and rhubarb.

Next to ornamental trees and shrubs, the most important nursery crops are fruit plants, followed by bulb crops. The most important single plant grown for outdoor cultivation is the rose. The type of nursery plants grown depends on location; in general (in the Northern Hemisphere) the northern areas provide deciduous and coniferous evergreens, whereas the southern nurseries provide tender broad-leaved evergreens.

The nursery industry includes wholesale, retail, and mail-order operations. The typical wholesale nursery specializes in relatively few crops and supplies only retail nurseries or florists. The wholesale nursery deals largely in plant propagation, selling young seedlings and rooted cuttings, known as "lining out" stock, of woody material to the retail nursery. The retail nursery then cares for the plants until growth is complete. Many nurseries also execute the design of the planting in addition to furnishing the plants.

Bulb crops. The bulb crops include plants such as the tulip, hyacinth, narcissus, iris, daylily, and dahlia. Included also are nonhardy bulbs used as potted plants indoors and summer outdoor plantings such as amaryllises, anemones, various tuberous begonias, caladiums, cannas, dahlias, freesias, gladioli, tigerflowers, and others. Hardy bulbs, those that will survive when left in the soil over winter, include various crocuses, snowdrops, lilies, daffodils, and tulips.

Tulips

Many bulb crops are of ancient Old World origin, introduced into horticulture long ago and subjected to selection and crossing through the years to yield many modern cultivars. One of the most popular is the tulip. Tulips are widely grown in gardens as botanical species but are especially prized in select forms of the garden tulip (which arose from crosses between thousands of cultivars representing several species). Garden tulips are roughly grouped as early tulips, breeder's tulips, cottage tulips, Darwin tulips, lily-flowered tulips, triumph tulips, Mendel tulips, parrot tulips, and others. The garden tulips seem to have been developed first in Turkey but were spread throughout Europe and were adopted enthusiastically by the Dutch. The Netherlands has been the centre of tulip breeding ever since the 18th century, when interest in the tulip was so intense that single bulbs of a select type were sometimes valued at thousands of dollars. The collapse of the "tulipmania" left economic scars for decades. The Netherlands remains today the chief source of tulip bulbs planted in Europe and in North America. The Netherlands has also specialized in the production of related bulbs in the lily family and provides hyacinth, narcissus, crocus, and others. The Dutch finance extensive promotion of their bulbs to support their market. Years of meticulous growing are required to yield a commercial tulip bulb from seed. Thorough soil preparation, high fertility, constant weeding, and careful record keeping are part of the intensive production, which requires much hand labour. Bulbs sent to market meet specifications as to size and quality, which assure at least one year's bloom even if the bulb is supplied nothing more than warmth and moisture. The inflorescence (flowering) is already initiated and the necessary food stored in the bulb. Under less favourable maintenance than prevails in The Netherlands, a subsequent year's bloom may be smaller and less reliable; it is not surprising therefore that tulip-bulb merchants suggest discarding bulbs after one year and replanting with new bulbs to achieve maximum yield.

Herbaceous perennials. Garden perennials include a number of herbaceous species grown for their flowers or occasionally used as vegetative ground covers. Under favourable growing conditions the plants persist and increase year after year. The biggest drawback to perennials as compared with annuals is that they must be maintained throughout the growing season but have only a limited flowering period. Typical perennials are hollyhocks, columbines, bellflowers, chrysanthemums, delphiniums, pinks, coralbells, phlox, poppies, primroses, and speedwells.

Perennials are often produced and sold as a sideline to other nursery activities; some are sold through seed houses. Perennial production could be undertaken on a massive scale, with attendant economies, but the market is neither large enough nor predictable enough (except for the greenhouse growing of such cut flowers as chrysanthemums and carnations) to interest most growers.

Shrubs. Production of ornamental shrubs is the backbone of the nursery trade in Europe and the United States. The nursery business is about equally divided between the production of (1) coniferous evergreens such as yew, juniper, spruce, and pine; (2) broad-leaved evergreens such as rhododendron, camellia, holly, and boxwood; (3) deciduous plants such as forsythia, viburnum, berberis, privet, lilac, and clematis; and (4) roses.

Fields of specialization have evolved within the ornamental shrub industry. Some firms confine activity mostly to production of "lining out" stock, which must be tended several years before reaching salable size.

The field grower may, in turn, specialize in mass growing for the wholesale trade only. The field plantings are tended until they attain marketable size. Because of the time required to produce a marketable crop and because of rising labour costs, this phase of the nursery industry involves economic hazards. But wholesale growing escapes the high overhead of retail marketing in urban areas, and, although many growers do sell stock at the nursery, they generally avoid the expensive merchandising required of the typical urban-area garden centre. Growers are especially interested in laboursaving technology and are turning to herbicidal control of weeds and shortcut methods for transplanting.

There is a well-established trade in container-grown stock—that is, nursery stock grown in the container in which it is sold. This practice avoids transplanting and allows year-round sales of plant material.

Roses. The production of roses is probably the most specialized of all shrub growing; the grower often deals solely in rose plants. Most are bud-grafted onto rootstocks (typically *Rosa multiflora*). This is the only way to achieve rapid and economical increase of a new selection to meet market demands. Large-scale production of roses has tended to centre in areas where long growing seasons make rapid production possible.

Because the budding operation calls for skilled hand labour and because field maintenance is expensive, few economies can be practiced in the production of roses. But distribution techniques that do offer certain economies have been developed. These include covering the roses with coated paper or plastic bags instead of damp moss to retain humidity and applying a wax coating to stems of dormant stock to inhibit desiccation.

Trees. Ornamental shade trees are usually grown and marketed in conjunction with shrubs. The 20th-century migration of people in many countries to suburban areas, coupled with the construction of houses on cleared land, has made shade trees an increasingly important part of the nursery trade. As interest in shade and ornamental trees increased, creation of improved cultivars followed. There is still some activity in transplanting native trees from the woodlot, and some are still grown from genetically unselected seed or cuttings; but more and more, like roses and shrubs before them, trees are vegetatively propagated as named cultivars, and many are patented.

The design and planning of landscapes has become a distinct profession that in many cases is only incidentally horticultural. Landscape architecture in its broadest sense is concerned with all aspects of land use. As a horticulturist, the landscape architect uses plants along with other landscape materials—stone, mortar, wood—as elements of landscape design. Unlike the materials of the painter or sculptor, plants are not static but change seasonally and with time. The colour, form, texture, and line of plants are used as design elements in the landscape. Plant materials are also manipulated as functional materials to control erosion, as surface materials, and for enclosures to provide protection from sunlight and wind.

Landscape architecture originated in the design of great estates, and home landscape is still an integral part of landscape architecture. More recently, however, landscape architecture has begun to include larger developments such as urban and town planning, parks both formal and "wild," public buildings, industrial landscaping, and highway and roadside development. (See GARDEN AND LANDSCAPE DESIGN.)

Horticultural education and research

Scholarly works in horticulture appear continuously in scientific literature. Specific institutions devoted to horticultural research, however, go back to the beginning of the experiment-station system, the first being a private laboratory of John Bennet Lawes, with the later collaboration of Joseph Henry Gilbert, in Rothamsted, Eng. (1843). Horticultural education and research in the United States was given great impetus by Justin S. Morrill, a supporter of the Morrill Act (1862), which provided educational institutions in agricultural and mechanical arts for each state. State experimental stations and the federal experimental stations of the U.S. Department of Agriculture, with its centre at Beltsville, Md., carry out systematic research efforts in horticulture. Although much research is carried out on horticultural food crops, there has been an increasing emphasis on ornamentals. Horticultural research is also conducted by private companies among the seed industry, canning and processing firms, and private foundations and botanical gardens.

Horticultural education is an established part of professional agricultural education worldwide. Training in horticulture up to the Ph.D. degree is offered in universities. There are relatively few schools devoted to the training of gardeners and horticultural technicians in the United States, although a number of state universities have two-year programs in horticulture. Vocational horticultural training is more highly developed in Europe.

There are a great number of national and international societies devoted to horticulture. These include community organizations such as garden clubs, specialty organizations devoted to a particular plant or group of plants (*e.g.*, rose and orchid societies), scientific societies, and trade organizations. The first society devoted to horticulture originated in 1804 with the establishment in England of the Royal Horticultural Society. There are similar organizations in other European countries. The American Pomological Society, dedicated to the science and practice of fruit growing, was formed in 1848. The American Horticultural Society, established in 1945, is devoted largely to ornamentals. The American Society for Horticultural Science was established in 1903 and became perhaps the most widely known scientific society devoted to horticulture. The International Society for Horticultural Science, formed in 1959 with permanent headquarters in The Hague, sponsors international congresses every four years. Most societies and horticultural organizations publish periodicals.

There are thousands of publications in the world devoted to some aspect of horticulture. The scientific and technical horticultural literature since 1930 is abstracted in *Horticultural Abstracts*, prepared by the Commonwealth Bureau of Horticulture and Plantation Crops, East Malling, Kent, Eng. (J.J./Ro.P.)

BIBLIOGRAPHY

Gardening. *History:* Short histories may be found in many of the larger gardening encyclopaedias. Fuller accounts include CHRISTOPHER THACKER, *The History of Gardens* (1979, reprinted 1985); ANTHONY HUXLEY, *An Illustrated History of Gardening* (1978, reprinted 1983); MILES HADFIELD, *A History of British Gardening*, 3rd ed. (1979); and JOSEPHINE VON MIKLOS and EVELYN FIORE, *The History, the Beauty, the Riches of the Gardener's World* (1969).

Classics and good reading: WILLIAM ROBINSON, *Wild Garden*, 4th ed. (1894, reprinted 1977); GERTRUDE JEKYLL, *Wood and Garden* (1899, reprinted 1983), *Home and Garden* (1900, reprinted 1982), and *Colour Schemes for the Flower Garden* (1908, reprinted 1983); GERTRUDE JEKYLL and LAWRENCE WEAVER, *Gardens for Small Country Houses* (1912, reprinted 1981); HENRY N. ELLACOMBE, *In a Gloucestershire Garden* (1895, reprinted 1982), and *In My Vicarage Garden and Elsewhere* (1902); MARIA THERESA EARLE, *Pot-Pourri from a Surrey Garden* (1897); V. SACKVILLE-WEST, *V. Sackville-West's Garden Book*, ed. by PHILIPPA NICOLSON (1968, reprinted 1983); MARGERIE FISH, *We Made a Garden* (1956, reissued 1983), *An All the Year Garden* (1958), *Cottage Garden Flowers* (1961, reissued 1980), and *A Flower for Every Day* (1965, reissued 1981); KAREL ČAPEK, *Gardener's Year* (1931, reprinted 1984; originally published in Czech, 1929); CHRISTOPHER LLOYD, *The Well-Tempered Garden*, new rev. ed. (1985), *The Adventurous Gardener* (1983), and *The Well-Chosen Garden* (1984); and JOHN RAVEN, *A Botanist's Garden* (1971).

Reference sources: A wealth of information can be found in the multivolume series *The Time-Life Encyclopedia of Gardening*. DOUGLAS REID, *Botany for the Gardener* (1966), is an account of principles underlying gardening practices; and RUSSEL PAGE, *The Education of a Gardener* (1962, reissued 1985), is helpful in garden planning. See also D.J. EDWARDS, *Gardening Explained* (1969); and HUGH JOHNSON, *The Principles of Gardening* (1979). More substantial and detailed works include CHRISTOPHER BRISKELL (ed.), *The Royal Horticultural Society's Concise Encyclopedia of Gardening Techniques* (1983); and ROY HAY and PATRICK M. SYNGE, *The Dictionary of Garden Plants in Colour*, new ed. (1976).

Garden types: LANNING ROPER, *Successful Town Gardening* (1957); JUDITH BERRISFORD, *The Wild Garden* (1966); XENIA FIELD, *Town and Roof Gardens* (1967); and KENNETH A. BECKETT, DAVID CARR, and DAVID STEVENS, *The Contained Garden* (1982).

Houseplants. ROY HAY *et al.*, *The Dictionary of Indoor Plants in Colour* (1974, reissued 1983); ANTHONY HUXLEY (ed.), *The World Guide to House Plants* (1983); and THOMAS ROCHFORD and RICHARD GORER, *The Rochford Book of Houseplants*, rev. ed. (1973).

Horticulture. AMERICAN HORTICULTURAL SOCIETY, *North American Horticulture* (1982); LIBERTY HYDE BAILEY and ETHEL ZOE BAILEY, *Hortus Third* (1976); THOMAS H. EVERETT, *New York Botanical Garden Illustrated Encyclopedia of Horticulture*, 10 vol. (1980–82); JULES JANICK, *Horticultural Science*, 3rd ed. (1979), and *Plant Science: An Introduction to World Crops*, 3rd ed. (1981).

Gastronomy

Gastronomy is the art of selecting, preparing, serving, and enjoying fine food. Anthelme Brillat-Savarin, the celebrated French aphorist and gastronomic authority of the late 18th and early 19th centuries, called gastronomy "the intelligent knowledge of whatever concerns man's nourishment."

Through the ages gastronomy has proved to be a stronger cultural force among the peoples of the world than linguistic or other influences. Today, the world may be divided into definite gastronomic regions, areas where distinctive cuisines prevail and common culinary methods are practiced.

Rice is the staple in most of Southeast Asia. The distinctive feature of the cooking of India and Indonesia is the generous and imaginative use of spices to lend an added zest to foods. Olive oil is the common denominator of the Mediterranean cuisines. Northern Europe and North America use a variety of cooking fats, among them butter, cream, lard, and goose and chicken fats, but the common gastronomic denominator throughout most of these lands is wheat, the basic crop. In Latin America corn (maize) is the staple and is used in a wide variety of forms.

This article covers the history of gastronomy from ancient civilizations through Greece and Rome, the Middle Ages, and the Italian Renaissance; the development of the great cuisines of France and of China; and the leading regional and national cuisines of the world.

The article is divided into the following sections:

HISTORY OF GASTRONOMY IN THE WEST

The first significant step toward the development of gastronomy was the use of fire by primitive man to cook his food, which gave rise to the first meals as families gathered around the fire to share the foods they had cooked. Prehistoric cave paintings such as those in Les Trois Frères in Ariège, in southern France, depict these early gastronomic events.

In the ancient civilizations of Assyria, Babylonia, Persia, and Egypt, the selection, preparation, service, and enjoyment of food were practiced on an elaborate scale. In the Book of Daniel the Bible relates the story of how Belshazzar, the king of the Chaldeans, "made a great feast to a thousand of his lords, and drank wine before the thousand." He then commanded gold and silver vessels to be brought, and he and his wives, princes, and concubines drank wine and praised gods of gold, silver, brass, iron, wood, and stone.

Greece and Rome. In ancient Greece, the Athenians believed that mealtime afforded an opportunity to nourish the spirit as well as the body. They reclined on couches while eating and accompanied their repasts with music, poetry, and dancing. The Greeks provided a philosophical basis for good living, Epicureanism. It held that pleasure was the main purpose of life; but pleasure was not intended to imply the self-indulgence that it connotes today. The Epicureans believed that pleasure could best be achieved by practicing self-restraint and indulging as few desires as possible. Today the epicure is defined as one who is "endowed with sensitive and discriminating tastes in food and wine."

Epicurean restraint and Roman excesses

The ancient Greeks practiced moderation in all things, but the Romans were known for their excesses. Ordinary citizens subsisted on barley or wheat porridge, fish, and ground pine nuts (edible pine seeds), but the Roman emperors and wealthy aristocrats gorged themselves on a staggering variety of foods. They staged lavish banquets where as many as 100 different kinds of fish were served, as well as mountainous quantities of beef, pork, veal, lamb, wild boar, venison, ostrich, duck, and peacock. They ordered ice and snow hauled down from the Alps to refrigerate their perishable foods, and they dispatched emissaries to outposts of the Roman Empire in search of exotic delicacies. Mushrooms were gathered in France, and the Roman author Juvenal, writing in the late first and early 2nd century AD, describes a dinner at his patron's house where mullet from Corsica and lampreys from Sicily were served.

Yet, whereas the Romans placed great value on exotic delicacies, they were not gastronomes in the true sense of the word. The term implies a sensitivity and discrimination that they lacked. The unbridled appetites of the Roman emperors and nobles often carried them to wild extremes. The emperor Caligula drank pearls that had been dissolved in vinegar. Maximus reportedly consumed 60 pounds of meat in a day, and Albinus was alleged to have eaten 300 figs, 100 peaches, 10 melons, and vast quantities of other foods all at a single sitting. Lucullus was an immensely wealthy man who entertained so lavishly that his name became a symbol both for extravagance and for culinary excellence.

The vulgarity and ostentation of Roman banquets were satirized by Petronius in the *Satyricon*, written in the 1st century AD. A former slave named Trimalchio entertains at a gargantuan feast at which the guests are treated to one outlandish spectacle after the other. A donkey is brought in on a tray, encircled with silver dishes bearing dormice that have been dipped in honey. A huge sow is carved and live thrushes fly up from the platter. A chef cuts open the belly of a roast pig, and out pour blood sausages and blood puddings.

The European Middle Ages. Through most of the Middle Ages banquets and feasts were characterized by their crudity and extravagance. In more affluent households in Gaul, for example, huge quarters of beef, mutton, and pork were served up whole before the guests. Wild game was frequently served, including wild boar, hedgehog, roebuck, crane, heron, and peacock. Meats and other foods were cooked over fires on spits and in caldrons positioned close by the tables. Diners seated themselves on bundles of straw and gobbled huge quantities of food, sometimes drawing knives from sheaths in their sword scabbards to saw off huge chunks of meat.

The first French cookbook of importance

The Frankish emperor Charlemagne introduced a touch of refinement. He decorated the walls of his banquet halls with ivy. Floors were strewn with flowers. Tables were laid with silver and gold utensils, but food was coarse, and menus offered little variety. The subtle seasonings and sauces that later were to characterize French cooking were still unknown. An insight into the relative crudeness of the French culinary art during the Middle Ages is provided by *Le Viander* (*c.* 1375), the first French cookbook of importance. It was written by Guillaume Tirel, more familiarly known as Taillevent, who served as chef to King Charles VI. Like the Romans, he used bread as the thickener for his sauces, instead of flour (which has been used for the past two centuries). He relied heavily on spices—such as ginger, cinnamon, cloves, and nutmeg (a Moorish influence via Spain and Italy). His menus consisted mostly of soups, meats, and poultry, which were so heavily seasoned that the taste of the food was largely obscured. French cooking ultimately would be distinguished by the subtlety of its seasonings and sauces and by the imaginative blend of textures and flavours. But in Taillevent's day the principal object of cooking was to disguise the flavour of the food (which, because of the lack of refrigeration, frequently was tainted) rather than to complement it.

The Italian Renaissance. The turning point in the development of gastronomic excellence was the Italian Renaissance. By the beginning of the 15th century wealthy merchants in Italy were dining in elegant style. In place of the crude slabs of meat served elsewhere, they savoured delicacies such as mushrooms, garlic, truffles, and tournedos (thick slices of beef fillets) and pasta dishes such as lasagne or ravioli. In wealthier households such delicacies were served in sumptuous style. When Vincenzo I, the duke of Mantua, celebrated his wedding feast in 1581, a guest reported that there were

> 100 ladies beautiful beyond measure and most richly garbed . . . [and there,] on a handsome sideboard, was visible a perspective of divers cups, carafes and goblets, and such beautiful vessels of Venetian glass as I think would defy description . . .

The duke's wedding feast lasted three hours. Tables were laid with delicately embroidered cloths. The guest reported that

> the first service from the sideboard was large salads decked out with various fantasies such as animals made of citron, castles of turnips, high walls of lemons; and variegated with slices of ham, mullet roes, herrings, tunny, anchovies, capers, olives, caviar, together with candied flowers and other preserves.

The duke's guest also reported that the tables were tastefully and imaginatively decorated.

> There were three large statues in marzipan. One was the horse of Campidoglio come to life, the second Hercules with the lion, and the third a unicorn with its horn in the dragon's mouth.

The bounty of these tables was almost beyond belief.

> The table was filled with many other things—jellies, blancmanges in half relief, spiced hard-bake, royal wafers, Milanese biscuits, pine kernels, minced meat, salami, cakes of pistachio nuts, sweet almond twists, flaky pastries . . . Indian turkey hens stuffed and roasted on the spit, marinated pullets, fresh grapes, strawberries strewn with sugar, wild cherries, and asparagus cooked in butter in various ways.

DEVELOPMENT OF FRENCH GRANDE CUISINE

Catherine de Médicis' Florentine cooks

The Italian influence on France. Italy has been called the mother of the Western cuisines, and perhaps its greatest contribution was its influence on France. The crucial event was the arrival of Catherine de Médicis in France in the 16th century. The great-granddaughter of Lorenzo the Magnificent, Catherine married the young man who later was to become Henry II of France. She brought with her a retinue of Florentine cooks who were schooled in the subtleties of Renaissance cooking—in preparing such elegant dishes as aspics, sweetbreads, artichoke hearts, truffles, liver crépinettes, quenelles of poultry, macaroons, ice cream, and zabagliones. Catherine also introduced a new elegance and refinement to the French table. Although, during Charlemagne's reign, ladies had been admitted to the royal table on special occasions, it was during Catherine's regime that this became the rule and not the exception. Tables were decorated with silver objects fashioned by Benvenuto Cellini. Guests sipped wine from fine Venetian crystal and ate off beautiful glazed dishes. An observer reported that

> the Court of Catherine de' Médici was a veritable earthly paradise and a school for all the chivalry and flower of France. Ladies shone there like stars in the sky on a fine night.

The French view of cooking

Catherine's cousin, Marie de Médicis, who married Henry IV of France, also advanced the culinary arts. An important new cookbook appeared in her time. It was called *Le Cuisinier françois* (1652) and was written by La Varenne, an outstanding chef, who is believed to have learned to cook in Marie de Médicis' kitchens. La Varenne's cookbook was the first to present recipes in alphabetical order, and the book included the first instructions for vegetable cooking. By now spices were no longer used to disguise the taste of food. Truffles and mushrooms provided subtle accents for meats, and roasts were served in their own juices to retain their flavours. A basic point of French gastronomy was established; the purpose of cooking and of using seasoning and spices was to bring out the natural flavours of foods—to enhance rather than disguise their flavour. In keeping with this important principle, La Varenne cooked fish in a fumet, or stock made with the cooked fish trimmings (head, tail, and bones). The heavy sauces using bread as a thickener were discarded in favour of the roux, which is made of flour and butter or another animal fat.

Contributions of the Sun King. La Varenne's cookbook was a gastronomic landmark, but a long time was to pass before the French cuisine would achieve its modern forms. In pre-Revolutionary France extravagance and ostentation were the hallmarks of gastronomy. Perhaps the most extravagant Frenchman of the time was the Sun King, Louis XIV, who, with members and guests of his court, wined and dined in unparalleled splendour at his palace at Versailles. There the kitchens were some distance from the King's quarters; the food was prepared by a staff of more than 300 people and was carried to the royal quarters by a procession headed by two archers, the lord steward, and other notables. As the cry "the King's meat" proclaimed their progress, an assemblage laden with baskets of knives, forks, spoons, toothpicks, seasonings, and spices solemnly made their way to the King's quarters. Before the King dined, tasters sampled the food to make certain it had not been poisoned. The King himself was such a prodigious eater that members of the court and other dignitaries considered it a privilege merely to stand by and watch him devour his food. His sister-in-law reported that at one meal he ate

> four plates of different soups, an entire pheasant, a partridge, a large plateful of salad, mutton cut up in its juice with garlic, two good pieces of ham, a plateful of cakes, and fruits and jams.

Table protocol and use of the fork

Louis XIV is remembered principally for his extravagance, but he was genuinely interested in the culinary arts. He established a new protocol for the table; dishes were served in a definite order instead of being placed on the table all at once without any thought to complementary dishes. The fork came to be widely used in France during his reign, and the manufacture of fine French porcelain was begun. The King himself hired a lawyer-agronomist, La Quintinie, to supervise the gardens at Versailles and was intensely interested in the fruits and vegetables—strawberries, asparagus, peas, and melons—that were grown there. He paid special honour to members of his kitchen staff, conferring the title of officer on his cooks.

Triumph of French grande cuisine. During the reigns of Louis XV and Louis XVI, culinary methods were refined and a new order and logic were introduced into the French cuisine. Brillat-Savarin noted that in the reign of Louis XV

> there was generally established more orderliness of meals, more cleanliness and elegance, and those refinements of service, which having increased steadily to our own time....

By the time of the Revolution the interest in the culinary arts had intensified to the point where Brillat-Savarin could report:

> The ranks of every profession concerned with the sale or preparation of food, including cooks, caterers, confectioners, pastry cooks, provision merchants and the like, have multiplied in ever-increasing proportions... New Professions have arisen; that, for example, of the pastry cook—in his domain are biscuits, macaroons, fancy cakes, meringues... The art of preserving has also become a profession in itself, whereby we are enabled to enjoy, at all times of the year, things naturally peculiar to one or other season... French cookery has annexed dishes of foreign extraction... A wide variety of vessels, utensils and accessories of every sort has been invented, so that foreigners coming to Paris find many objects on the table the very names of which they know not, nor dare to ask their use.

The Revolution and the rise of the restaurant

The Revolution changed almost every aspect of French life: political, economic, social, and gastronomic as well. In pre-Revolutionary days the country's leading chefs performed their art in wealthy aristocratic households. When the Revolution was over those who had survived the guillotine and remained in the country found employment in restaurants. The restaurant became the principal arena for the development of the French cuisine, and henceforth French gastronomy was to be carried forward by a succession of talented chefs, men whose culinary genius has not been matched in any other land. From the long roll of great French chefs, a select company have made a lasting contribution to gastronomy.

The great French chefs. The first, and in many ways the most important of all French chefs, was Marie-Antoine Carême, who has been called the Architect of the French Cuisine. As the French novelist and gastronome Alexandre Dumas *père* related the story, Carême was born shortly before the Revolution, the 16th child of an impoverished stonemason; when he was only 11, his father took him to the gates of Paris one evening, fed him supper at a tavern, and abandoned him in the street. Fortunately for gastronomy, Carême found his way to an eating house, where he was put to work in the kitchen. Later he moved to a fine pastry shop where he learned not only to cook but also to read and draw.

Carême, the "Architect of the French Cuisine"

Carême was an architect at heart. He liked to spend time in strolling about Paris, admiring the great classic buildings, and he fell into the habit of visiting the Bibliothèque Royale, where he spent long hours studying prints and engravings of the great architectural masterpieces of Greece, Rome, and Egypt. He designed massive, elaborate table decorations called *pièces montées* ("mounted pieces") as an outlet for his passionate interest in architecture. In an age of Neoclassicism his tables were embellished with replicas of classic temples, rotundas, and bridges constructed with spun sugar, glue, wax, and pastry dough. Each of these objects was fashioned with an architect's precision, for Carême considered confectionery to be "architecture's main branch," and he spent many months executing these designs, rendering every detail with great exactness.

Carême was employed by the French foreign minister, Charles-Maurice de Talleyrand, who was not only a clever diplomat but a gastronome of distinction. Talleyrand believed that a fine table was the best setting for diplomatic manoeuvring. Following his service with Talleyrand, Carême practiced his art for a succession of kings and nobles. He catered a series of feasts for Tsar Alexander of Russia, was *chef de cuisine* to England's Prince Regent (who later became George IV), and finally was employed by Baroness Rothschild in Paris.

Today Carême's monumental table displays seem ostentatious almost beyond belief, but he lived in an opulent age that was obsessed with classical architecture and literature. To appreciate him fully, one must put him in the context of his time. Before Carême, the French cuisine was largely a jumble of dishes; little concern was given to textures and flavours and compatibility of dishes. Carême brought a new logic to the cuisine. "I want order and taste," he said. "A well displayed meal is enhanced one hundred per cent in my eyes." Every detail of his meals was planned and executed with the greatest of care. Colours were carefully matched, and textures and flavours carefully balanced. Even the table displays, mammoth as they were, were designed and carried out with an architect's precision.

Carême's voluminous cookbooks, *L'Art de la cuisine au dix-neuvième siècle* (1833) and *Le Pâtissier royal parisien* (1815), included hundreds of recipes, menus for every day in the year, a history of French cooking, sketches for Carême's monumental *pièces montées*, instructions for garnishes, decorations, and tips on marketing and organizing the kitchen.

Montagné and Escoffier

After Carême, the two men who probably had the greatest impact on French gastronomy and that of the world at large were Prosper Montagné and Georges-Auguste Escoffier. Montagné was one of the great French chefs of all time, and he achieved a secure place in gastronomic history by creating *Larousse Gastronomique* (1938), the basic encyclopaedia of French gastronomy. As a young man, while serving as an assistant chef at the Grand Hotel in Monte-Carlo, he came to the conclusion that all *pièces montées*, as well as superfluous garnitures and decorations, should be discarded. This was a drastic step, and Montagné's call might have gone unheeded had it not been brought to the attention of Escoffier. Escoffier was unimpressed at first. But his friend and literary collaborator Philéas Gilbert (also an outstanding chef) persuaded him that Montagné was right. Escoffier became a zealot for culinary reform, insisting on refining and modifying nearly every aspect of the cuisine. He simplified food decorations, greatly shortened the menus, accelerated the service, and organized teams of cooks to divide and share tasks in order to prepare the food more expertly and efficiently.

All of this progress was greatly facilitated by the introduction of the Russian table service around 1860. Before then, the service *à la française* was used. Under that method the meal was divided into three sections or services. All of the dishes of each service were brought in from the kitchen and arrayed on the table at once. Then, when this service was finished, all of the dishes of the next service were brought in. The first service consisted of everything from soups to roasts, and hot dishes often cooled before they could be eaten. The second service comprised cold roasts and vegetables, and the third was the desserts. Under the Russian table service, which was popularized by the great chef Félix Urbain-Dubois, each guest was served each course individually, while the food was at its best.

Escoffier invented scores of new dishes. One was *poularde Derby*, roast chicken with rice, truffles, and foie gras stuffing, garnished with truffles and foie gras. Other better known Escoffier inventions were *pêche Melba*, a dessert made with peaches, and Melba toast, tributes to the Australian soprano Nellie Melba.

In naming some of his culinary creations after friends and celebrities Escoffier was following a well-established tradition. *Tournedos Rossini*, the tender slices of the heart of the fillet of beef, topped by foie gras and truffles, was named after the celebrated Italian composer. The composer Giuseppe Verdi and the actress Sarah Bernhardt were among those who were similarly honoured. Many famous dishes have taken their names from the chefs who invented them—*Sole Dugléré*, for example, was named after the chef Adolphe Dugléré, who presided at the Café Anglais in Paris in the middle of the 19th century. French dishes have also been named after their dominant colours, such as *carmen* or *cardinale*, which refers to a pinkish-reddish hue. Great events have also given dishes their names: chicken Marengo, for example, was named after the battle in 1800 in which Napoleon defeated the Austrians.

Escoffier created a cold dish called chicken Jeannette. It was named after a ship that was crushed by icebergs. Escoffier's creation was stuffed breast of chicken, and, in honour of the ill-fated ship, he served it on top of a ship carved out of ice.

The *grande cuisine* of France is the only structured and organized system of gastronomy in the world. Many dishes are interrelated, and their names contain clues as to their ingredients. For example, soups are broken down into consommés (clear soups), potages (thick soups), crèmes (cream soups), and veloutés (made with a white sauce). Within each of these categories there are sub-categories, depending upon the base used, the thickening agent, the garniture, the flavouring spice, herb, or alcohol, and other considerations.

Escoffier's fame today rests mostly on the cookbooks he wrote—*Le Livre des menus* (1924), *Ma Cuisine* (1934), and *Le Guide culinaire* (1921), written in collaboration with Gilbert—in which he codified the French cuisine in its modern form, setting down thousands of menus and clarifying the principles of French gastronomy. With the great hotel entrepreneur César Ritz, he established a string of the world's most luxurious hostelries in Paris, Rome, Madrid, New York, Budapest, Montreal, Philadelphia, and Pittsburgh.

In the late 1950s a group of young French chefs led by Paul Bocuse, Michel Guérard, the Troisgros brothers Jean and Pierre, and Alain Chapel invented a free-form style of cooking (named *nouvelle cuisine* by the French restaurant critics Henri Gault and Christian Millau). Their style disregarded the codification of Escoffier and replaced it with a philosophy rather than a structured system of rules, creating not a school but an anti-school, in reaction to the French *grande cuisine*. The basic characteristics of *nouvelle cuisine* included the replacement of the thickening of sauces with reductions of stocks and cooking liquids; the serving of novel combinations in very small quantities artistically arranged on large plates; a return to the importance of purchasing of food; and infinite attention to texture and detail.

At its best, *nouvelle cuisine* produced dishes that avoided rich sauces and lengthy cooking times, and its creative and inventive practitioners aroused interest and excitement in gastronomy generally and in restaurants specifically. One of its most important results has been the recognition of the chef as a creative artist.

Use of wines and sauces. French gastronomy is distinguished not only by the genius of its chefs but also by well-established culinary practice. One of these practices is the use of fine wines, such as those produced in Bordeaux and Burgundy, as accompaniments to good food. The proper choice of wines—according to vintages, vineyards, shippers—is an indispensable part of French gastronomy.

In the preparation of food, the hallmark of French gastronomy is the delicate sauces that are used to enhance the flavours and textures. Sauces are prepared with stocks, or *fonds de cuisine,* "the foundations of cooking." These stocks are made by simmering meats, bones, poultry or fish trimmings, vegetables, and herbs in water to distill the essence of their flavours.

The basic sauces

There are literally hundreds of French sauces, but among the more familiar ones are the families of white sauces, brown sauces, and tomato sauces, the mayonnaise family, and the hollandaise family. White sauces, which are served with poultry, fish, veal, or vegetables, are prepared by making a white roux, a mixture of butter and flour, which is cooked and stirred to smoothness. Béchamel sauce is prepared by adding milk and seasoning to this thickening agent. Sauce velouté is made by mixing a fish, poultry, or veal stock with the roux. A broad variety of sauces is derived from these basic white sauces. Sauce mornay is simply sauce béchamel with grated cheese and seasonings. Sauce suprême is béchamel with cream. Sauce normande is prepared by mixing a fish velouté with tarragon and white wine or vermouth.

Brown sauces, which are served with red meats, chicken, turkey, veal, or game, are prepared by simmering a meat stock for many hours and then thickening it with a brown roux, a mixture of butter and flour cooked until it turns brown. Among the better known brown sauces are sauce ragout, which is flavoured with bone trimmings or giblets; sauce diable, which is seasoned with lots of pepper; sauce piquant, a brown sauce with pickles and capers, and sauce Robert, which is seasoned with mustard.

The hollandaise family is another important branch of French sauces. Hollandaise is closely related to mayonnaise. It is prepared by delicately flavouring warmed egg yolks with lemon juice and then carefully stirring in melted butter until the mixture achieves a creamy, yellow thickness. Sauce mousseline is made by adding whipped cream to hollandaise, while sauce béarnaise also has an egg and butter base with tarragon, shallots, wine, vinegar, and pepper. Sauce vin blanc is made by adding a white-wine fish stock to the basic hollandaise.

GASTRONOMY IN CHINA

The gastronomic tradition

Apart from the French cuisine, the highest expression of the gastronomic art is generally regarded to be that of the Chinese. It is no accident that China and France should have produced the world's most distinctive and respected cuisines. Both countries were naturally blessed with an abundance and rich variety of raw ingredients. In each of these countries gastronomy traditionally commanded great interest and respect. The intellectual, artistic, political, and financial leaders of China and France traditionally attached great importance to good eating. It has already been noted how this worked in the case of the Bourbon kings of France and with statesmen of such eminence as Talleyrand. In ancient China the preparation and service of food played an important part in court rituals. The first act of many emperors was to appoint a court chef, and once they were on the job these chefs strove mightily to outdo each other.

In ancient China, hunting and foraging supplied much of the food. Wild game, such as deer, elk, boar, muntjac (a small deer), wolf, quail, and pheasant, was eaten, along with beef, mutton, and pork. Vegetables such as royal fern, smartweed, and the leafy thistle (*Sonchus*) were picked off the land. Meats were preserved by salt-curing, pounding with spices, or fermenting in wine. To provide a contrast in flavours the meat was fried in the fat of a different animal.

As Chinese agriculture developed, styles of food were determined to a great degree by the natural resources available in certain parts of the country, thus the vastly different manners of cooking and the development of distinctive regional cuisines of China. As a more varied fare began to emerge, tastes grew more refined. By the time of Confucius (551–479 BC) gastronomes of considerable sophistication had appeared on the scene. Confucius wrote of one of these fastidious eaters,

> For him the rice could never be white enough. When it was not cooked right, he would not eat. When the food was not in season, he would not eat. When the meat was not cut correctly, he would not eat. When the food was not served with the proper sauce, he would not eat.

Emergence of a cuisine. Like all other forms of haute cuisine, classic Chinese cooking is the product of an affluent society. By the 2nd century AD the Chinese court had achieved great splendour, and the complaint was heard that idle noblemen were lounging about all day, feasting on smoked meats and roasts.

By the 10th or 11th century a distinctive cuisine had begun to emerge, one that was developed with great attention to detail. It was to reach its zenith in the Ch'ing dynasty (1644–1911/12). This cuisine was a unique blend of simplicity and elegance. The object of cooking and the preparation of food was to extract from each ingredient its unique and most enjoyable quality.

As in the case of the French cuisine, the hors d'oeuvre set the tone of the meal. "The hors d'oeuvre must look neat," say the Chinese gastronomic authorities Tsuifeng Lin and her daughter Hsiang Ju Lin.

> They are best served in matched dishes, each containing one item. Many people like to garnish the dishes with parsley and vegetables cut in the shape of birds, fish, bats, etc., or even to make baskets of flowers from food. These are all acceptable if kept under control, and if the rest of the meal is served in the same florid style. The worst offense would be to start with a florid display of food and then suddenly change style midway . . .

Common foods and traditions. The theory of balancing *fan* (grains and rice) with *ts'ai* (vegetables and meat) is one of the factors that distinguish Chinese gastronomy from that of all other nations. This refined proportion of harmony and symmetry of ingredients was practiced whenever possible in households throughout the ages and is not limited to formal or high cuisine or to meals served on special occasions.

In addition to taste that pleases (a most elemental requirement in China), astrological, geographical, and personal characteristics had to satisfy the complex system of the *yin–yang* balance of hot and cold, based on Taoist perception of the cosmic equilibrium. According to this theory, every foodstuff possesses an inherent humour; thus, consuming foods and beverages at proper and complementary temperatures can adjust the possible deviation of the normal state of the two intertwining forces.

Certain foods and culinary traditions are prevalent throughout most of the country. Rice is the staple except in the north, where wheat flour takes its place. Fish is extremely important in all regions. Pork, chicken, and duck are widely consumed, as well as large quantities of such vegetables as mushrooms, bamboo shoots, water chestnuts, and bean sprouts. The Chinese season their dishes with monosodium glutamate and soybean sauce, which takes the place of salt. Another distinctive feature of Chinese cooking is the varied and highly imaginative use of fat, which is prepared in many different ways and achieves the quality of a true delicacy in the hands of a talented Chinese cook. The Chinese take tea with their meals, whether green or fermented. Jasmine tea is served with flowers and leaves in small-handled cups.

The great Chinese schools. Traditionally, China is divided into five gastronomic regions, three of which are characterized by the great schools of Chinese cooking, Peking, Szechwan, and Chekiang-Kiangsu. The other two regions, Fukien and Kwangtung, are of lesser importance from a gastronomic point of view.

Peking. Peking is the land of fried bean curd and water chestnuts. Among foods traditionally sold by street vendors are steamed bread and watermelon seeds. Vendors also dispensed buns called *paotse* that were stuffed with pork and pork fat, and *chiaotse,* or crescents, cylindrical rolls filled with garlic, cabbage, pork, scallions, and monosodium glutamate. Wheat cakes wrapped around a filling of scallions and garlic, and noodles with minced pork sauce are also traditional Peking specialties. But the greatest of all delicacies of this region is of course the Peking duck. This elaborate, world-renowned dish requires lengthy preparation and is served in three separate courses. In its preparation, the skin is first puffed out from the duck by introducing air between the skin and the flesh. The duck is then hung out to dry for at least 24 hours, preferably in a stiff, cold breeze. This pulls the skin away from the meat. Then the duck is roasted until the skin is crisp and brown. The skin is removed, painted with Hoisin sauce (a sweet, spicy sauce made of soybeans), and served inside the folds of a bun as the first course. The duck meat is carved from the bones and carefully cut into slivers. Sautéed onions, ginger, and peppers are added to the duck meat and cooked with bean sprouts or bamboo slivers. This forms the second course. The third course is a soup. The duck bones are crushed and then water, ginger, and onion are added to make a broth. The mixture is boiled, then drained, and the residue is cooked with cabbage and sugar until the cabbage is tender.

The preparation of Peking duck

Szechwan. The cooking of Szechwan in central China is distinguished by the use of hot peppers, which are indigenous to the region. The peppers lend an immediate sensation of fiery hotness to the food, but, once this initial reaction passes, a mingled flavour of sweet, sour, salty, fragrant, and bitter asserts itself. Fried pork slices, for example, are cooked with onions, ginger, red pepper, and soy sauce to achieve this aromatic hotness.

Chekiang and Kiangsu. The provinces of Chekiang and Kiangsu feature a broad variety of fish—shad, mullet, perch, and prawns. Minced chicken and bean-curd slivers are also specialties of these provinces. Foods are often arranged in pretty floral patterns before serving.

Fukien. Fukien, which lies farther south, features shredded fish, shredded pork, and *popia,* or thin bean-curd crepes filled with pork, scallions, bamboo shoots, prawns, and snow peas.

Kwangtung. To Americans perhaps, the most familiar form of Chinese cooking is that of Kwangtung, for Canton lies within this coastal province. Mushrooms, sparrows, wild ducks, snails, snakes, eels, oysters, frogs, turtles, and winkles are among the many exotic ingredients of the province. More familiar to Westerners are such Cantonese specialties as egg roll, egg foo yung, and roast pork.

The Cantonese specialties

OTHER CUISINES OF THE WORLD

Japanese. Nowhere has greater care and imagination been given to the presentation of food than in Japan. The delicacy and exquisiteness of Japanese table arrangements are matched only by the fragile beauty of Japanese painting.

Traditionally the Japanese bride received as many as 50 different kinds of dishes as wedding gifts, and she might use a dozen at one meal. She would devote the most painstaking attention to the angle at which a sprig of green vegetable was propped against a lump of crabmeat, or the way a fish was garnished. Meals were served in many small dishes, but the total amounts offered each diner were large.

The waters around Japan abound with fish and shellfish, and Japanese seafood is regarded by many gourmets as the finest in the world. Fish is eaten raw (*sashimi*), broiled, fried in deep fat (*tempura*), or salted and broiled (*shioyaki*). The popular *tempura* method of deep frying food was learned from Portuguese traders who came to Japan in the 16th century. Rice has been the staple; it traditionally accompanied every meal; but in the late 20th century wheat products such as bread have become common, especially as an accompaniment to Western-style food. *Sushi,* or vinegared rice, is served in stylized portions with a variety of accompaniments, including mushrooms, squid, fish, shrimp, and caviar.

The Japanese like clear soups, garnished with eggs, vegetables, or seafood. The thicker "miso" soups are flavoured with fermented soybean paste. Japanese vegetables include bamboo shoots, snow peas, eggplant, mushrooms, and potatoes. The popular sukiyaki consists of beef and vegetables simmered in soy sauce. Pork or chicken may be substituted for the beef. Saké, a fermented beverage made from rice or other grain, is a popular drink, and tea is taken with all meals and at virtually all hours of the day.

The Japanese tea ceremony, or chanoyu, is a highly formalized ritual dating back to the 13th century. The tea is meticulously prepared and is accompanied by a variety of delicate seasonal dishes. Every aspect of the ceremony—the setting, the flavours and textures of foods, the colours and shapes of the containers, even the conversation—is carefully calculated to achieve the most harmonious and satisfying effect.

The Japanese tea ceremony

An outgrowth of the tea ceremony is the *kaiseki,* the *grande cuisine* of Japan; it is the highest form of Japanese dining and perhaps comes as close to dining as an art form as any in the entire world of gastronomy. The food served in *kaiseki* is selected according to the changing seasons and is presented through a series of small dishes with an artful simplicity that brings out the unique tastes of ordinary foods from nearby mountains and sea. Perhaps the key to the composition of the *kaiseki* meal lies in the word *aishoh:* "compatibility."

Indian. Spices are a distinctive feature of the cooking of India and Indonesia. In India, every good cook prepares a curry—a mixture of such fragrant powdered spices as cardamom, cinnamon, cloves, cumin, nutmeg, and turmeric. The spice blend is kept in a jar in the kitchen and is used to season all sorts of foods.

The Hindus of India have developed what is perhaps the world's greatest vegetarian cuisine. They use cereals, pulses (lentils, peas, and beans), and rice with great imagination to produce a widely varied but generally meatless cuisine.

Indian cooks prepare delicious chutneys, highly seasoned vegetables and fruits used as side dishes that must be fresh to be fully appreciated. They also make little delicacies such as *idli*s, cakes of rice and lentils that are cooked by steaming; *pakora*s, vegetables fried in chickpea batter; and *jalebi*s, pretzel-like tidbits made by soaking a deep-fried batter of wheat and chickpea flour in a sweet syrup. *Raytas*, yogurt with fruits or vegetables, are another favourite. Other specialties include *biryānī,* a family of complicated rice dishes cooked with meats or shrimp; *samosa,* a flaky, stuffed, deep-fried pastry; *korma,* lamb curry made with a thick sauce using crushed nuts and yogurt; *masala,* the dry or wet base for curry; and a great variety of breads and hot wafers, including *naan, pappadam, parāṭhā*s, and *chapātī*s.

In southern India and especially in the historical region of Telingana, or Andhra, the food is seasoned with fresh chili peppers and can be fiery hot. Lamb is the most important meat served in northern India. It is prepared in hundreds of different ways as kabobs, curries, roasts, and in rice dishes. In pre-independence days the Mughal cuisine there ranked among the most lavish in the world. The Mughal cuisine developed during the Muslim empire of the great Mughal kingdom. It is based, mostly because of religious and geographic limitations, on lamb. The preparations are mostly roasted, barbecued dishes, also kabobs and the so-called dry curries, versus the stew-type cooking of the south.

In India festivals and holidays are marked by feasting and revelry. Among the more prominent festivals are Ōṇam, a rice harvest celebration; Dīwālī, which marks the beginning of the Hindu New Year; Dashera, which marks the triumph of the good prince Rama over evil; and Holī, the festival of lights, which honours Lord Krishna, an incarnation of the god Vishnu. Feasting and the offering of food to gods and friends are a highlight of these festivals.

The Pacific and Southeast Asia. The cuisine of the Pacific and Southeast Asia is a fascinating melange of raw ingredients, methods, and dishes with a strong influence of the Chinese cuisine. The most important ingredients to tie together this vast area are the coconut, which is used in every one of these countries; rice, which is the basic food everywhere except in the Philippines; and native spices and herbs, especially the omnipresent ginger and chili. The skillful use of condiments and relishes by its inventive cooks makes each of these countries a gastronomically individual entity.

Coconut, rice, spices, and herbs

A Hawaiian staple is the taro bulb, which is the main ingredient for many dishes of the famous luau feasts. Taro may be chopped and steamed alone, or mixed with other ingredients, often wrapped in ti leaves. Poi is made by peeling and cooking the taro root and then mashing it into a paste.

Another famous delicacy is *lomi lomi,* a fresh salmon that is massaged by hand to break down its tissues and remove the salt. Chunks of the fish are mixed with onion and tomatoes. Besides the stone-baked pig, which is always a part of the luau, and several other local specialties, the Hawaiians adapted a number of Chinese, Japanese, Korean, and Indian dishes, together with a great many standard U.S. dishes.

Indonesia consists of several thousand islands, yet its cuisine is almost unified by the use of coconut. It is employed as a vegetable, main course, ingredient, cooking fat, relish, fruit, and even beverage in the popular *tjendol* throughout the islands. Although 300 years of Dutch occupation, a sizable Chinese population, and Portuguese merchants had a very strong influence on the islands' cooking style, Indonesia still can boast of a unique cuisine. Because rice (*nasi*) is the most important part of the meal, all other preparations are actually served to surround and enhance the rice itself. The Dutch themselves created *rijsttafel* (literally, "rice table"), which formalized into an almost endless procession of beautifully arranged, carefully organized dishes, ranging from sweet to sour, from mild to very spicy, from cold to hot.

Each guest was given two plates and was served a long succession of excellent dishes. On one plate a meat or fish preparation would be served, and the other would be filled with rice. The entire meal, with all its courses, took from two to three hours. Since Indonesia gained independence, the *rijsttafel* has been replaced by the *prasmanan,* a lengthy, buffet-style meal also featuring scores of dishes. *Rijsttafel* became popular in The Netherlands, however, and could be ordered in many restaurants, particularly in Amsterdam.

One of the nationally popular preparations is *nasi goreng,* which originates in China's fried-rice concept. In the Indonesian version, however, most of the meats, vegetables, and garnishes surround the pile of fried rice and only the diner mixes them while eating it, allowing many fascinating taste and texture combinations.

Although many dishes are common to all areas, each region has its own specialties and style of cooking. The West Javanese cooking is rather mild and tends to be much simpler than that of Central Java, which favours very hot, rich, and sweet flavours. East Java, on the other hand, is the place where the spicing becomes very complex and subtle, and the Balinese enjoy many of the dishes forbidden to the Muslim population. For instance, the Bali Hindu religion allows the eating of pork, and *saté babi,* the little skewers of charcoal-grilled pork bits, is one of the more interesting of their preparations.

One of the most essential elements of an Indonesian meal is the *sambal*s. These are spicy-hot condiments that are served separately to be mixed with the various foods to make them as "fiery" as the individual desires. *Krupuk,* the deep-fried shrimp wafers, also originated in Indonesia before turning up in other nations' cuisines. Few Indonesian meals are served without *gado-gado,* an interesting melange of cooked and raw vegetables and bean cake with a sauce made of peanuts, coconut, and spices. Sumatra and Malaysia absorbed much of the Arab and Indian culinary influences. *Rendang,* for instance, is a beef stew that absorbs a large amount of coconut milk, using the same technique as some of the so-called dry curries of India. *Gulai* is this area's favourite version of liquid-type curry so common in India.

The Philippine food is much simpler than many of the other Pacific and Southeast Asian cuisines. Although the four centuries of Spanish domination brought considerable influence to this part of the world, Philippine cuisine does have some specialties that can be called its own. Perhaps most typical of these is the fish paste called *bagoong* and the liquid flavouring sauce *patis.* Both are based on fermented seafood and, depending on the area or the household, their variety is almost limitless. Generally speaking, a sour-salty taste is the single most characteristic taste of the Philippines.

Perhaps the strongest Chinese influence can be detected in Vietnam, which was dominated or ruled by China through most of the 1st millennium AD. The degree of influence is discernible even in the manner of eating. For instance, this is the only country in the entire area of the Pacific and Southeast Asia where the food is eaten with chopsticks. *Nuoc mam,* a flavouring sauce, is used in many dishes, and, although it is related to the Philippine *patis,* it really is a specifically Vietnamese flavour, based again on fermented salted fish and spices. Almost every nation's southern inhabitants prefer their food spicier than those in the northern region, and Vietnam is no exception. The tie-in perhaps between the two regions of Vietnam is the use of fish, which is the most important part of the daily diet. The French occupation in Vietnam mostly contributed to the level of the gastronomy of the upper classes, without influencing very much of the average housewife's cooking.

One of the most complex and structured cuisines of the entire area is the cuisine of Thailand. The fact that the Thai lived for much of their history in comparative peace and political independence had beneficial influence on their gastronomy, together with the fact that, just as in China and France, the ruling classes were actively interested in gastronomy. Because the Thai have basically the same ingredients to work with as the Indonesians, Malaysians, or Indians, the categories of the Thai cooking repertoire are not dissimilar, but the subtleties and complexities of flavour and texture are often superior. For instance, *nam prik,* the spicy Thai condiment, has even more varieties than the Indonesian *sambal*s do, with many more ideas employed in their combinations. *Kaeng* is a liquid stew (or perhaps soup-stew) to be mixed with rice. It is very strongly related to the liquid curries, but again the repertoire of *kaeng*s is infinitely larger than almost any other food family in Southeast Asia. Within the formalized gastronomy, the Chinese and Indian influences blend in with such artistry that the emerging cuisine of Thailand is truly its own.

Middle Eastern. Eggplant, olives, and yogurt are widely eaten in all Middle Eastern countries. Chickpeas are toasted or ground. Lamb is the staple meat throughout the region. One of the most characteristic elements of the cuisines of the Middle East is the offering of an almost unlimited array of small hot and cold appetizers. These are called *mazza* (Arabic), *mezethakia* (Greek), or *mezelicuri* (Romanian), and their ingredients and preparation have developed over the centuries as a result of the confluence of many cultures.

Culinary influence of the old Ottoman Empire

The Turkish influence is still dominant in the countries of the old Ottoman Empire: Turkey, Greece, Bulgaria, and other parts of the Balkan region. Vine leaves stuffed with rice and meat are popular. They are called *dolma* in Turkey. *Börek,* a turnover filled with meat or cheese, is another favourite. *Şişkebabi* (shish kebab), skewered mutton or lamb, is enjoyed in all these countries, as is *kofte,* a lamb patty. Yogurt dishes and a sweet known as halvah are commonly found. A favourite dessert is baklava, a rich pastry filled with nuts and layered with honey or syrup. (Baklava was brought by Turkish invaders in the 16th century to Hungary, where it became strudel.)

The Arab states of the Middle East and North Africa share many fine dishes. Among these is the hotly seasoned eggplant dip called *bābā qhanūj.* Other dishes common to the Arab countries include *hummus bī tahīnah,* chickpeas with a sesame paste; *tabbūlah,* a salad of onions, chopped tomatoes, radishes, parsley, and mint; and *kibbi,* a ground mixture of wheat and lamb.

African. The great indigenous dish of North Africa is couscous—steamed wheat or semolina grains—served with meats, poultry, and vegetables piled on top of and around the grain. Another is *brik* (related to the Turkish *börek*), a deep-fried pastry turnover stuffed with fish or meat and a whole egg. A sampler from this continent would include, in East Africa, peanut (groundnut) soup and beef and cassava stew, cooked with coconut, chilies, and coriander; in West Africa, fish *imojo,* a fish and seafood salad; in Ethiopia, *yetemola cheguara,* steamed stuffed whole tripe. South Africans prepare a cinnamon- and clove-flavoured stew, called lamb and pumpkin *bredi,* and date and onion salad. Angola has a yellow coconut pudding in which the colour comes from the predominance of eggs among the ingredients.

Spanish and Portuguese. Spain and Portugal have much in common from a culinary point of view. Olive oil is the cooking fat of both countries. Cod is widely used. The *cocido,* a heavy stew of boiled chicken, meats, and vegetables, is Spain's national dish. In Portugal it is called the *cozido.*

But the two countries also have their own distinctive dishes, which vary greatly from one region to the next. The paella is perhaps Spain's best known dish. It is a colourful combination of rice, chicken, pork, clams, mussels, shrimp, peppers, sausages, and peas. The home of the paella is Valencia, but the dish varies from one province to the next. Another regional specialty is the *zarzuela de mariscos,* a Catalan seafood medley, a stew of fish, lobster, shrimp, scallops, clams, ham, almonds, white wine, and saffron. Fish is popular throughout Spain, especially cod, hake, and red snapper. Many people consider the Basque-style cooking (*à la Vasca*) the best in Spain. It is a surprisingly sophisticated cuisine for one based on ancient shepherds' cooking. *Tapas* are appetizers served in Spanish bars, and often there are several dozen varieties from which to choose. *Jamón serrano,* a mountain-cured ham; chorizo sausages; gazpacho, a cold soup made of pureed vegetables and generally quite spicy; and meat pies called empanadas are some of the highlights of the quite remarkable cuisine of Spain.

The Portuguese kitchen produces somewhat spicier and richer foods, favouring hearty soups, marinated seafoods, braised meats, and such spices as cumin and coriander. The entire Iberian Peninsula is strongly influenced by the honey-almond paste, figs, and dates typically used by the Moors.

Italian. The Italians are especially fond of pasta *asciutta* (an unending variety of dried noodles), the huge assortment of hot and cold appetizers known as antipasti; sausage and salami; *gelati e granite,* ice creams and ices; and *caffè espresso,* coffee made by forcing steam through the coffee grounds.

Italy, like France or China, has many culinary regions, but basically the north's staple is rice and butter and the south lives on pasta and cooks with olive oil. Cooking techniques are less important than the quality of the raw ingredients.

Bologna's rich cooking is perhaps the best of the northern cuisine with its famed *tagliatelle, tortellini,* and other freshly made noodle preparations, egg pastas, sausages, and complex main courses. Piedmont supplies many of the finest chefs to the luxury restaurants around the world. Its local white truffles and Fontina cheese are the base for their *fonduta,* the famous hot melted cheese casserole eaten with bread bits.

Lombardy cooks exclusively with butter, replacing the pasta with rice and cornmeal polenta, and blends successfully the cooking style of several of the northern provinces. Genoese cooking's most characteristic flavour comes from the use of basil leaves pounded into a sauce called *pesto* together with cheese, garlic, pine nuts, and olive oil. Florence is famous for its *Chianina* beef cattle that provide the meat for its *bistecca alla Fiorentina.*

Alla Romana-type cooking produces the best gnocchi, *calamaretti* (baby squid), *abbacchio* (young lamb, usually roasted with rosemary), and vegetable preparations. The cooking of Naples represents the best gastronomy of southern Italy with the use of pasta, crusty white bread, robust tomato sauces, mozzarella, and other types of cheese. Generally speaking, the availability of some of the finest vegetables and fruits of Europe and of fine seafood, and the array and liberal use of fresh herbs, create the best moments of the Italian gastronomy.

Austro-Hungarian. The gastronomic regions of the world existing today do not conform to geographic or political divisions. A good example is the Austro-Hungarian cuisine, a culinary entity the boundaries of which have not been discernible on maps since the end of World War I. This gastronomical region comprises the old Austro-Hungarian Empire. It includes Austria and Hungary, as well as parts of Romania and other areas of the Balkan region, the Czech Republic, and Slovakia. The people of these countries live in different political, economic, and social systems and speak different languages, but their culinary heritage remains as a link between them. *Gulyás,* or goulash, is prepared in varying forms in all of these countries. Wiener schnitzel (breaded veal cutlets, named for the city of Wien, or Vienna) is eaten throughout the area. Coffeehouses are popular throughout this part of the world. In Austria a traveller will encounter *nockerl,* a small dumpling; in Hungary, he may eat *nokedli;* in the Czech Republic and in Slovakia he will find a similar food under the name *noky,* in Serbia under the name *nokla.* In Hungary the *nokedli* would accompany *pörkölt,* a stew made by browning onion in lard and adding paprika, or a *paprikás,* similar to the above, but with the addition of sweet or sour cream. A dessert would be *Rigó Jancsi,* a chocolate square glazed with chocolate and filled with chocolate mousse. Cakes, tortes, and desserts are the glory of this cuisine. Prune dumplings, strudels, the coffee ring called *gugelhupf,* and the Dobos torte, a caramel-topped cake filled with chocolate-cocoa cream, are enjoyed. And one of the greatest glories of Vienna's old empire is Sacher torte, a chocolate sponge cake with a touch of apricot jam, iced with a bittersweet chocolate.

Each section of this gastronomic region developed its own specialties. The people of the Czech-Slovak region, for instance, cook *játernice* (a sausage made with innards), bake *koláĉ* (coffee cake), and make *povidla* (prune jam). Austria offers *Backhendl* (breaded-fried spring chicken), *Tafelspitz* (the boiled beef considered the finest, a cut near the tail of the steer), Linzer torte (a ground-almond and jam lattice cake), *buchteln* (jelly buns), and *Palatschinken* (thin pancakes with filling). In the Yugoslav region popular items are the *čevapčići* (charcoal-grilled meatballs), *gibanica* (cheesecake), *lonac* (a Bosnian meat and vegetable stew), and *šumadija* (a tea made with plum brandy and caramelized sugar).

Slavic. Russia is the mother country of the Slavic cuisine, another culinary entity that does not exist on the map. This cuisine comprises the former Soviet Union, Poland, Albania, and parts of the Yugoslav region and Bulgaria. A Russian-speaking traveller through this region might find it difficult to make himself understood, but he could order the familiar borsch, a beet or cabbage soup, wherever he went. He might dine on blintzes (stuffed pancakes) or *zrazy* (stuffed fried fish or seafood). He could enjoy beef stroganoff, beef cooked with onions in sour cream, or a seafood pie called *rakov*. Wherever he went, vodka would be the most popular drink.

The Russians developed *zakusky*, their equivalent of the French hors d'oeuvres. *Potage Bagration* (cream of veal with asparagus tips) is also part of the French *grande cuisine* together with many other dishes the French chefs learned in the Russian court kitchens. Interesting specialties are the *botvinya* (green vegetable soup with a fish base), *solyanka* (cucumber soup), *pelemeni* (Siberian meat dumplings, boiled, fried, and served with sour cream), *kasha* (buckwheat porridge), *holubtsi* (Ukrainian stuffed cabbage), *bitki* (meatballs or fish balls with strong spices), *paskha* (cottage-cream cheesecake with candied fruits made in a pyramid shape for Easter), and *babka* (a round coffee cake).

German. The emphasis of the cuisine of Germany and its neighbours is on "hearty" foods—roast meats, dumplings, fish dishes, cream sauces, puddings, and rich desserts. The Germans eat sauerbraten, a marinated pot roast with a sweet-and-sour sauce, the earthy hasenpfeffer (hare stew), *Königsberger Klopse* (a fancy meatball), *badischer Hecht* (a sour cream baked pike), and *Schweinebraten mit Pflaumen und Apfeln* (roast pork with prune and apple stuffing). Other favourite foods are sausages; sauerkraut (fermented cabbage); dumplings; thick soups made from potatoes, peas, or lentils; herring; and roast meats, or braten. Popular desserts include puddings, fruit pancakes or dumplings, egg custards, jellies topped with whipped cream, the medieval invention marzipan (an almond paste confection), lebkuchen (a kind of gingerbread), and *Baumkuchen* (the "tree-cake" baked on a special horizontal spit).

Scandinavian. Fish is a mainstay of the Scandinavian diet. It is prepared in many different ways; a favourite appetizer is *gravlax*, salmon marinated in salt and dill and accompanied by a mustard sauce. Swedish pancakes are popular and are served with lingonberries or fruit preserves.

The sumptuous bread-and-butter table

Sweden's great contribution to international eating is the smorgasbord, literally a "bread-and-butter table" but actually a sumptuous feast of three courses. The first course is herring—filleted, pickled, baked, jellied, stewed, or prepared in many other different ways. Cold meats constitute the second course, whereas the third course is made up of Swedish meatballs and other hot dishes.

Danish open sandwiches, called *smørrebrød*, became popular all over the world. Among the many fine dishes in these northern European countries are *nyponsoppe* (a Swedish soup prepared with rose hips, almonds, and whipped cream), *vorshmack* (ground meat, herring, and onion cooked Finnish style), "Jansson's Temptation" (a Swedish potato and anchovy casserole), *frikadeller* (a Danish mixed ground meat hamburger, sautéed in butter), *kalakukko* (a Finnish bird-shaped pie stuffed with fish), *sandkage* (Danish sand cake), and *krumkage* (a Norwegian Christmas cookie). Aquavit is the favourite grain or potato spirit of many in Scandinavia.

British. Favourites among the English are roast beef with Yorkshire pudding, an accompaniment similar in texture to a popover; steak and kidney pie; and veal and ham pie. Fish is served often—plaice (a type of flounder), haddock, mackerel, and smoked kipper—and especially popular are fish-and-chips (deep-fried fish and potatoes). Jellies, jams, marmalade, hot cross buns, crumpets, and scones are served frequently with tea.

Traditional fare in the British Isles would include beef tea (a beef extract), whitebait (miniature fish, fried and eaten as snacks), boxty (Irish potato pancakes), brawn (aspic made with pork bits), cockaleekie (Scottish hen and leek soup), bubble and squeak (chopped, fried leftover meat and vegetables), angels on horseback (grilled oysters wrapped in bacon), kedgeree (a casserole of smoked fish, rice, and eggs), shepherd's pie (ground lamb and beef with onion and topped with mashed potatoes), crumpets, banbury cake (a spiced flat cake made with dried fruits), fool (a fruit custard), and syllabub (a dessert made with whipped cream, lemon, wine, and sugar).

Latin American. Corn (maize) is the culinary common denominator of much of Latin America. Ground into meal, it is used in Mexico to prepare the corn pancakes known as tortillas. Tortillas provide a variety of other Mexican specialties. Enchiladas are tortillas dipped in sauce, then rolled up with a filling of pork or chicken and baked or broiled. *Tostadas* are tortillas fried crisp and sprinkled with onion, chili peppers, grated cheese, or meat. *Quesadillas* are tortillas folded over a filling of meat, beans, cheese, or vegetables. Corn is also used in tamales, which are steamed, filled corn husks.

Chili peppers are widely used to season Latin American dishes. Brazil's national dish, the *feijoada completa,* consists of a bed of rice with black beans, sausages, beef tongue, spareribs, and dried beef, sprinkled with toasted manioc meal, and garnished with orange slices. In Argentina, empanadas are beef-filled turnovers. A favourite Latin American dessert is flan, or caramel custard.

Quick-service and convenience foods

North American. While North American cities such as New York, New Orleans, San Francisco, Chicago, and Montreal have produced many excellent restaurants and hotels, the unique American contribution to gastronomy has been quick-service and convenience foods. The first cafeteria came into being in San Francisco during the Gold Rush of 1849. Automated cafeterias were later introduced in New York and Philadelphia.

The United States is a culinary melting pot. In New York City and many other metropolitan areas, one can find almost any kind of food. Outside the great cities, American food at one time had a distinctive regional character. New England was famous for its clam and lobster dishes, its New England boiled dinner, and its red flannel hash. The South had its fried chicken, barbecued meats, and corn breads. The Far West prided itself on its Dungeness crab, abalone, fish, and shellfish. As a result of easy transportation of fresh, packaged, and frozen foods, once strictly regional dishes have become popular countrywide. A "new" American cooking, combining inventive simplicity and eclectic venturesomeness, offers a challenge to the bastions of European gastronomy.

BIBLIOGRAPHY. PROSPER MONTAGNÉ, *Larousse Gastronomique*. new ed. (1988; originally published in French, rev. and corrected ed., 1967), is the most authoritative contemporary encyclopaedia of food, wine, and cookery, from prehistoric stages to the modern day. It may be supplemented by WAVERLEY ROOT, *Food: An Authoritative and Visual History and Dictionary of the Foods of the World* (1980); JEAN-FRANÇOIS REVEL, *Culture and Cuisine: A Journey Through the History of Food* (1982; originally published in French, 1979); MARGARET VISSER, *The Rituals of Dinner: The Origins, Evolution, Eccentricities, and Meaning of Table Manners* (1991); and MAGUELONNE TOUSSAINT-SAMAT, *A History of Food* (1992; orginally published in French, 1987).

Early works, many of them now classics, include JEAN ANTHELME BRILLAT-SAVARIN, *The Physiology of Taste; or, Meditations on Transcendental Gastronomy*, trans. by M.F.K. FISHER (1949, reissued as *M.F.K. Fisher's Translation of the Physiology of Taste*, 1986; originally published in French, 2 vol., 1826), a classic that set the stage for thinking about dining as an experience and a form of art, in a translation by a distinguished American essayist on gastronomy; ANDRÉ L. SIMON, *A Concise Encyclopaedia of Gastronomy*, 9 vol. (1939–46, reissued in 1 vol., 1981). a complete history; URBAIN DUBOIS and ÉMILE BERNARD, *La Cuisine classique*, 2 vol. (1856), the finest expression of the Golden Age of the French *grande cuisine;* ALEXIS SOYER, *The Pantropheon* (1853, reprinted 1977), a world history of food preparation with many arbitrary but important observations on gastronomy; and ABRAHAM HAYWARD, *The Art of Dining*, ed. by CHARLES SAYLE (1899), on 18th- and 19th-century gastronomy, chefs, and related subjects. Chinese cuisine is treated in PEARL KONG CHEN, TIEN CHI CHEN, and ROSE Y.L. TSENG, *Everything You Want to Know About Chinese Cooking* (1983); and K.C. CHANG (ed.), *Food in Chinese Culture: Anthropological and Historical Perspectives* (1977).

(G.L./Ed.)

Gauss

Carl Friedrich Gauss, who, with Archimedes and Newton, ranks as one of the greatest mathematicians of all time, at an early age overturned the theories and methods of 18th-century mathematics and, following his own revolutionary theory of numbers, opened the way to a mid-19th-century rigorization of analysis. Although he contributed significantly to pure mathematics, he also made practical applications of importance for 20th-century astronomy, geodesy, and electromagnetism. His own dictum, "Mathematics, the queen of the sciences, and arithmetic, the queen of mathematics," aptly conveys his perception of the pivotal role of mathematics in science.

Born on April 30, 1777, in Brunswick, now in Germany, Gauss was the only son of poor parents. Impressed by his ability in mathematics and languages, his teachers and his devoted mother recommended him to the Duke of Brunswick, who granted him financial assistance to continue his education in secondary school and from 1795 to 1798 to study mathematics at the University of Göttingen. In 1799 he obtained his doctorate in absentia from the university at Helmstedt. The subject of his dissertation was a proof of the fundamental theorem of algebra—which was proven only partially before Gauss—which states that every algebraic equation with complex coefficients has complex solutions; moreover, Gauss skillfully formulated and proved this theorem without the use of complex numbers.

Contributions to number theory

At age 24 he published the *Disquisitiones Arithmeticae,* one of the most brilliant achievements in the history of mathematics, in which he formulated systematic and widely influential concepts and methods of number theory—dealing with relationships and properties of integers (−2, −1, 0, +1, +2, . . .)—which, for him, was of paramount importance in mathematics. He dealt extensively with the theory of congruent numbers—(*i.e.,* those numbers that have the same remainder when they are divided by another number (for example, 7 and 9 are congruent modulo the number 2 since there is a remainder of 1 when each is divided by 2); he gave the first proof of the law of quadratic reciprocity, which has to do with the quadratic residues (a is called quadratic residue with respect to b, if there is an integer x such that when a is divided by b, the remainder is the same as x^2 divided by b); and he applied this law to special cases of equations in which he was able to bring together algebraic, arithmetic, and geometric ideas. Using number theory, for example, Gauss proposed an algebraic solution to the geometric problem of constructing a regular polygon that has n sides. Euclid had shown that regular polygons, with 3, 4, 5, and 15 sides and those the sides of which result from doubling the above could be constructed geometrically with compass and ruler. No progress had been made in this subject since then. Gauss developed a criterion based on number theory by which it can be decided whether a regular polygon with any given number of corners can be geometrically constructed: these include, for example, the regular polygon with 17 sides, which he inscribed within a circle using only compass and ruler, the first such discovery since the time of Euclid.

By courtesy of the Archiv der Georg-August-Universitat, Gottingen, Germany

Gauss, oil painting by C.A. Jensen (1792–1870). In the Archiv der Georg-August-Universität, Göttingen, Germany.

This work on number theory contributed to the modern arithmetical theory of algebraic numbers—that is, to the solution of algebraic equations—in which Gauss introduced the first step—that is, the arithmetic of all complex numbers $a + b\sqrt{-1}$, in which a and b are integers. The complex numbers $a + b\sqrt{-1}$ had been introduced only intuitively before Gauss. In the *Disquisitiones Arithmeticae* Gauss did not hesitate to use complex numbers $a + b\sqrt{-1}$, in which a and b are real numbers. In 1831 (published 1832) he gave a detailed explanation of how an exact theory of complex numbers can be developed with the aid of representation in the x, y plane.

Astronomical and geodetic research

In 1801 Gauss had the opportunity to apply his superior computational skills in a dramatic way and, by so doing, to express gratitude to the Duke for assisting him in obtaining an education. On the first day of the year, a body, subsequently identified as an asteroid and named Ceres, was discovered as it seemed to approach the Sun. Astronomers had been unable to calculate its orbit, although they could observe it for 40 days until lost from view. After only three observations Gauss developed a technique for calculating its orbital components with such accuracy that several astronomers late in 1801 and early in 1802 were able to locate Ceres again without difficulty. As part of his technique, Gauss used his method of least squares, developed about 1794, a method by which the best estimated value is derived from the minimum sums of squared differences in a particular computation. This achievement in astronomy won Gauss prompt recognition. His methods, which he described in his book, in 1809, *Theoria Motus Corporum Coelestium,* are still in use today, and only a few modifications have been required to adapt his methods for modern computers. He had similar success with the asteroid Pallas, for which he refined his calculations to take into account the perturbations of its orbit by planets.

The Duke continued to finance Gauss's research so generously that in 1803 he was able to decline an offer of a professorship in St. Petersburg, where he was by then a corresponding member of the Academy of Sciences. In 1807 he became professor of astronomy and director of the new observatory at the University of Göttingen, where he remained for the rest of his life. His first wife died in 1809, after a marriage of four years and soon after the birth of their third child. From his second marriage (1810–31) were born two sons and a daughter.

About 1820 Gauss turned his attention to geodesy—the mathematical determination of the shape and size of the Earth's surface—to which he devoted much time in theoretical studies and field work. To increase the accuracy of surveying he invented the heliotrope, an instrument by which sunlight could be utilized to secure more accurate measurements. By introducing what is now known as the Gaussian error curve, he showed how probability could be represented by a bell-shaped curve, commonly called the normal curve of variation, which is basic to descriptions of statistically distributed data. He also was interested in determining the shape of the Earth by actual geodetic measurements, which led him back to pure theory. Using data from these measurements, he developed a theory of curved surfaces by which characteristics of a surface could be found solely by measuring the lengths of the curves that lie on the surface. This "intrinsic-surface theory" inspired one of his students, Bernhard Riemann, to develop a general intrinsic geometry of spaces with three or more dimensions. It was the subject of Riemann's inaugural lecture at Göttingen in 1854 and is said to have agitated Gauss. About 60 years later Riemann's ideas formed the mathematical basis for Einstein's general theory of relativity.

Contributions to non-Euclidean geometry and to physics

Gauss was one of the first to doubt that Euclidean geometry was inherent in nature and thought. Euclid was the first to build a systematic geometry. Certain basic ideas in his model are called axioms; they were the points of departure from which his entire system was constructed through pure logic. Of these, the parallel axiom played a prominent role from the beginning. According to this axiom, only one line can be drawn parallel to a given line through any point not on the given line. From this axiom soon arose the supposition that it can be deduced out of the other axioms and thus can be omitted from the system of axioms. All proofs of it, however, contained errors, and Gauss was one of the first to realize how there might be a geometry in which the parallel axiom does not apply. Gradually he came to the revolutionary conclusion that there is indeed such a geometry that is internally consistent and free of contradiction. Because it ran counter to contemporary views, he feared publication (see GEOMETRY: *Non-Euclidean geometry*).

When a Hungarian, János Bolyai, and a Russian, Nikolay Lobachevsky, independently published a non-Euclidean geometry about 1830, Gauss announced that he had made the same conclusions approximately 30 years before. Neither did he publish his work on special complex functions, perhaps because he was unable to derive them from more general principles. Thus, this theory had to be reconstructed by other mathematicians from his calculations in work extending over several decades after his death.

Closely related to his interest in gravitation and magnetism was his published paper in 1840 on real analysis. This paper became the starting point for the modern theory of potential. It is probably the only work he did that failed to meet his own high standards. Only at the beginning of the 20th century was it possible for mathematicians to develop potential theory anew, on the basis of different principles or by finding the conditions under which Gauss's conclusions are completely correct.

About 1830, principles of extremals (maximum and minimum quantities) began to assume a substantial role in his mathematical investigations of physical problems, such as the conditions in which a fluid remains at rest. In his treatment of capillary action, he devised mathematical formulations that took into account the mutual actions of all the particles in a fluid system, the force of gravity, and the interaction of its fluid particles and the particles of solid or fluid with which it is in contact. This work contributed to the development of the principle of the conservation of energy. From 1830, Gauss worked closely with the physicist Wilhelm Weber. Because of their interest in terrestrial magnetism, they organized a worldwide system of stations for systematic observations. The most important result of their work in electromagnetism was the development, by other workers, of electric telegraphy. Because their finances were limited, their experiments were on a small scale; Gauss was rather frightened at the thought of worldwide communication.

Personality

Gauss was deeply religious, aristocratic in bearing, and conservative. He remained aloof from the progressive political currents of his time. In Gauss, apparent contrasts were combined in an effective harmony. A brilliant arithmetician with a phenomenal memory for numbers, he was at once a profound theoretician and an outstanding practical mathematician. Teaching was his only aversion, and, thus, he had only a few students. Instead, he effected the development of mathematics through his publications, about 155 titles, to which he devoted the greatest care. Three principles guided his work: "Pauca, sed matura" ("Few, but ripe"), his favourite saying; the motto "Ut nihil amplius desiderandum relictum sit" ("That nothing further remains to be done"); and his requirement of utmost rigour. It is evident from his posthumous works that there are extensive and important papers that he never published because, in his opinion, they did not satisfy one of these principles. He pursued a research topic in mathematics only when he might anticipate meaningful relationships of ideas and results that were commendable because of their elegance or generality.

The golden anniversary of the granting of the doctorate to Gauss was celebrated in 1849. For this event, he prepared a new edition of his earlier proofs of the fundamental theorem of algebra, which, because of his declining health, was his last publication. The honour that gave him the greatest joy, however, was the bestowal of honorary citizenship on him by the city of Göttingen. On the basis of his outstanding research in mathematics, astronomy, geodesy, and physics, he was elected as a fellow in many academies and learned societies. He declined numerous invitations of other universities to become a professor and remained on the faculty of the University of Göttingen until his death on February 23, 1855. Soon after his death, coins were struck in his honour. The title of *mathematicorum princeps* is a fitting tribute. (H.Re.)

BIBLIOGRAPHY. *Carl Friedrich Gauss Werke,* 12 vol. (1863–1933), presents Gauss's publications, posthumous works, part of his correspondence, and commentaries by the publishers. Biographies include W.K. BÜHLER, *Gauss* (1981); TORD HALL, *Carl Friedrich Gauss,* trans. from Swedish (1970); and W. SARTORIUS VON WALTERSHAUSEN, *Gauss, a Memorial* (1966; originally published in German, 1856), written by a friend as a nonmathematical account of Gauss's life. HANS REICHARDT (ed.), *C.F. Gauss Gedenkband anlässlich des 100. Todestages am 23. Februar 1955* (1957), contains essays on various aspects of Gauss's work, as well as facts on his life and activities; and GEORGE M. RASSIAS (ed.), *The Mathematical Heritage of C.F. Gauss* (1991), includes diverse essays covering theories and problems that Gauss initially set forth. (H.Re./Ed.)

The Principles of Genetics and Heredity

Heredity is the sum of all biological processes by which particular characteristics are transmitted from parents to their offspring. Among organisms that reproduce sexually, progeny are not exact duplicates of their parents but usually vary in many traits. Heredity and variation, two sides of the same coin, are the subject matter of the science of genetics. Genetics may be defined as the study of the way in which genes—the functional units of heritable material—operate and are transmitted from parents to offspring. Modern genetics also involves study of the mechanism of gene action; that is, the way in which the genetic material affects physiological reactions within the cell.

In many languages the same words are used for both the inheritance of biological traits and the inheritance of property. Biological and legal inheritances are, however, very different processes. Inherited objects are actually transferred from one owner to another. Inherited traits are not. Offspring inherit a genetic constitution from their parents. This hereditary endowment, the sum total of the genes that the individual has received from both parents, is called the genotype. The genotype must be contrasted to the phenotype, which is the organism's outward appearance: its bodily structures, physiological processes, behaviour, etc. Although the genotype determines the broad limits of the features an organism may develop, the features that actually develop—*i.e.,* the phenotype—depend upon complex interactions between genes and their environment. Since the environment, both internal and external, of an individual changes continuously, so does the phenotype. Thus the same individual shows different phenotypes in childhood, in adulthood, and in old age. The genotype, on the other hand, does not change during an individual's lifetime. In conducting genetic studies it is crucial to discover the degree to which the observable trait (the phenotype) is attributable to the pattern of genes in the cells (the genotype) and to what extent it arises from environmental influence.

The essence of heredity is the reproduction of the carriers of genetic information, the genes. As a result, biological organisms, including human beings, reproduce organisms resembling themselves; human children are always recognizably human and have phenotypes similar to those of their parents. On the other hand, since the offspring of sexually reproducing organisms receive varying combinations of genetic material from both parents, no two offspring (except for identical twins) have exactly the same genotype. This genetic diversity is always modified by an equally diverse environment, so the resulting phenotype is never exactly the same, even among identical twins.

Genetics is often called the core science of biology. This does not necessarily mean that genetics is the most fundamental among the biological disciplines. It implies only that genetics impinges upon almost every kind of study of life. Anthropology, medicine, biochemistry, physiology, psychology, ecology, systematics, comparative morphology, and paleontology all have intersections with genetics. Like so many basic, or "theoretical," sciences, genetics has many actual and potential practical applications. The understanding and control of hereditary disorders and the breeding of improved crops and livestock are just two such applications.

Knowledge of heredity dates to prehistoric times and has been applied to the breeding of plants and animals for centuries. Most of the mechanisms of heredity, however, remained a mystery until the 20th century. The pioneering work in elucidating the mechanisms of gene action took place even more recently, and the science of genetics is considered as yet in its infancy.

This article examines the discoveries that led to an understanding of heredity and discusses in detail the structure and function of the gene, mutation and other processes by which genetic information is altered, and the science of human genetics, applying some of the aspects mentioned earlier in the article specifically to human inheritance.

The article is divided into the following sections:

BASIC FEATURES OF HEREDITY

Early conceptions of heredity

Heredity was for a long time one of the most puzzling and mysterious phenomena of nature. This was so because the sex cells, which form the bridge across which heredity must pass between the generations, are usually invisible to the naked eye. Only after the invention of microscopes early in the 17th century, and the discovery of the sex cells, could the essentials of heredity be grasped. Before that time, Aristotle (4th century BC) speculated that the relative contributions of the female and the male parents were very unequal—the female was thought to supply what he called the "matter" and the male the "motion." The *Institutes of Manu,* composed in India between AD 100 and 300, consider the role of the female like that of the field and of the male like that of the seed; new bodies are formed "by the united operation of the seed and the field." In reality both parents transmit the heredity pattern equally, and, on the average, children resemble their mothers as much as they do their fathers. Nevertheless, the female and male sex cells may be very different in size and in structure; the mass of an egg cell is sometimes millions of times greater than that of a spermatozoon.

The ancient Babylonians knew that pollen from a male date palm tree must be applied to the pistils of a female tree to produce fruits. Rudolph Jacob Camerarius showed in 1694 that the same is true in corn (maize). Carolus Linnaeus in 1760 and Josef Gottlieb Kölreuter, in a series of works published from 1761 to 1798, described crosses of varieties and species of plants. They found that these hybrids were on the whole intermediate between the parents, although in some characteristics they may be closer to one and in others closer to the other parent. Kölreuter compared the offspring of reciprocal crosses—*i.e.,* of crosses of variety *A* functioning as a female to variety *B* as a male and the reverse, variety *B* as a female to *A* as a male. The hybrid progenies of these reciprocal crosses were usually alike, indicating that, contrary to the belief of Aristotle, the hereditary endowment of the progeny was derived equally from the female and the male parents. Many experiments on plant hybrids were made in the 1800s. These investigations revealed that hybrids were usually intermediate between the parents. They more or less incidentally recorded most of the facts that later led Gregor Mendel (see below) to formulate his celebrated rules and to found the theory of the gene. Apparently, none of Mendel's predecessors saw the significance of the data that were being accumulated. The general intermediacy of hybrids seemed to agree best with the belief that heredity was transmitted from the parents to offspring by "blood," and this belief was accepted by most 19th-century biologists, the evolutionist Charles Darwin being included among these.

Blood theory of heredity

The blood theory of heredity, if this notion can be dignified with such a name, is really a part of the folklore antedating scientific biology. It is implicit in such popular phrases as "half blood," "new blood," "blue blood." It does not mean that heredity is actually transmitted through the red liquid in blood vessels; the essential point is the belief that a parent transmits to each child all its characteristics and that the hereditary endowment of a child is an alloy, a blend of the endowments of its parents, grandparents, and more remote ancestors. This idea appeals to those who pride themselves on having a noble or remarkable "blood" line. It strikes a snag, however, when one observes that a child has some characteristics that are not present in either parent but are present in some other relatives or were present in more remote ancestors. Even more often one sees that brothers and sisters, though showing a family resemblance in some traits, are clearly different in others. How could the same parents transmit different "bloods" to each of their children?

Mendel disproved the blood theory. He showed (1) that heredity is transmitted through factors (now called genes) that do not blend but segregate, (2) that parents transmit only one-half of the genes they have to each child, and they transmit different sets of genes to different children, and (3) that, although brothers and sisters receive their heredities from the same parents, they do not receive the same heredities (an exception is identical twins). Mendel thus showed that, even if the eminence of some ancestor were entirely the reflection of his genes, it is quite likely that some of his descendants, especially the more remote ones, would not inherit these "good" genes at all. In sexually reproducing organisms, humans included, every individual has a unique hereditary endowment.

Lamarckism—a school of thought named for the 19th-century pioneer French evolutionist Jean-Baptiste, chevalier de Lamarck—assumed that characters acquired during an individual's life are inherited by his progeny, or, to put it in modern terms, that the modifications wrought by the environment in the phenotype are reflected in similar changes in the genotype. If this were so, the results of physical exercise would make exercise much easier or even dispensable in a person's offspring. Not only Lamarck but also other 19th-century biologists, including Darwin, accepted the inheritance of acquired traits. It was questioned by August Weismann, whose famous experiments in the late 1890s on amputation of tails in generations of mice showed that such modification resulted neither in disappearance nor even in shortening of the tails in the descendants. Weismann concluded that the hereditary endowment of the organism, which he called the germ plasm, is wholly separate and is protected against the influences emanating from the rest of the body, the somatoplasm or soma. The germ plasm–somatoplasm are related to the genotype–phenotype concepts, but they are not identical and should not be confused with them.

Superstitions regarding heredity

The noninheritance of acquired traits does not mean that the genes cannot be changed by environmental influences: X rays and other mutagens certainly do change them, and the genotype of a population can be altered by selection. It simply means that what is acquired by parents in their physique and their intellect is not inherited by their children. Related to these misconceptions are the beliefs in "prepotency"—that some individuals impress their heredities on their progenies more effectively than others—and in "prenatal influences" or "maternal impressions"—that the events experienced by a pregnant female are reflected in the constitution of the child to be born. How ancient these beliefs are is suggested in the Book of Genesis, in which Laban produced spotted or striped progeny in sheep by showing the pregnant ewes striped hazel rods. Another such belief is "telegony," which goes back to Aristotle; it alleged that the heredity of an individual is influenced not only by his father but also by males with whom the female may have mated and who have caused previous pregnancies. Even Darwin, as late as 1868, seriously discussed an alleged case of telegony: that of a mare that was mated to a zebra and subsequently to an Arabian stallion by whom the mare produced a foal with faint stripes on his legs. The simple explanation for this result is that such stripes occur naturally in some breeds of horses.

All these beliefs, from inheritance of acquired traits to telegony, must now be classed as superstitions. They do not stand up under experimental investigation and are incompatible with what is known about the mechanisms of heredity and about the remarkable and predictable properties of the genetic materials. Nevertheless, some people still cling to these beliefs. Some animal breeders take telegony seriously and do not regard as "pure bred" the individuals whose parents are admittedly "pure" but whose mothers had mated with males of other breeds. The agronomist Trofim Denisovich Lysenko was able for close to a quarter of a century, roughly between 1938 and 1963, to make his special brand of Lamarckism the official creed in the Soviet Union and to suppress most of the teaching and research in orthodox genetics. He and his partisans published hundreds of articles and books allegedly proving their contentions, which effectively deny the achievements of biology for at least a century. The Lysenkoists were officially discredited in 1964.

Mendelian genetics

DISCOVERY AND REDISCOVERY OF MENDEL'S LAWS

Gregor Mendel published his work in the proceedings of the local society of naturalists in Brünn, Austria (now Brno, Czechoslovakia), in 1866, but none of his contemporaries appreciated its significance. It was not until 1900, 16 years after Mendel's death, that his work was rediscovered independently by Hugo de Vries in Holland, Carl Erich Correns in Germany, and Erich Tschermak von Seysenegg in Austria. Like several investigators before him, Mendel experimented on hybrids of different varieties of a plant. Mendel investigated the common pea plant (*Pisum sativum*). His methods differed in two essential respects from those of his predecessors. First, instead of trying to describe the appearance of whole plants with all their characteristics, Mendel followed the inheritance of single, easily visible and distinguishable traits, such as round versus wrinkled seed, yellow versus green seed, purple versus white flowers, etc. Second, he made exact counts of the numbers of plants bearing this or that trait; it was from such quantitative data that he deduced the rules governing inheritance.

Experiments with pea plants

Since pea plants reproduce usually by self-pollination of their flowers, the varieties Mendel obtained from seedsmen were "pure"—*i.e.,* descended for several to many generations from plants with similar traits. Mendel crossed them by deliberately transferring the pollen of one variety to the pistils of another; the resulting first-generation hybrids, denoted by the symbol F_1, usually showed the traits of only one parent. For example, the crossing of yellow-seeded plants with green-seeded ones gave yellow seeds; the crossing of purple-flowered plants with white-flowered ones gave purple-flowered plants, etc. Traits such as the yellow-seed colour and the purple-flower colour Mendel called dominant; the green-seed colour and the white-flower colour he called recessive. It looked as if the yellow and purple "bloods" overcame or consumed the green and white "bloods."

That this was not so became evident when Mendel allowed the F_1 hybrid plants to self-pollinate and produce the second hybrid generation, F_2. Here both the dominant and the recessive traits reappeared, as pure and uncontaminated as they were in the original parents (generation P). Moreover, these traits now appeared in constant proportions: about 3/4 of the plants in the second generation showed the dominant trait and 1/4 showed the recessive, a 3 to 1 ratio. It can be seen in Table 1 that Mendel's actual counts were as close to the ideal ratio as one could expect allowing for the sampling deviations present in all statistical data.

Table 1: Pea Plants with Dominant and Recessive Characters Obtained by Mendel in the Second Generation of Hybrids

number dominant		number recessive		ratio
Round seed	5,474	Wrinkled seed	1,850	2.96 : 1
Yellow seed	6,022	Green seed	2,001	3.01 : 1
Purple flowers	705	White flowers	224	3.15 : 1
Tall plants	787	Short plants	277	2.84 : 1

Mendel concluded that the sex cells, the gametes, of the purple-flowered plants carried some factor that caused the progeny to develop purple flowers, and the gametes of the white-flowered variety had a variant factor that induced the development of white flowers. In 1909 the Danish biologist Wilhelm Ludvig Johannsen proposed to call these factors genes.

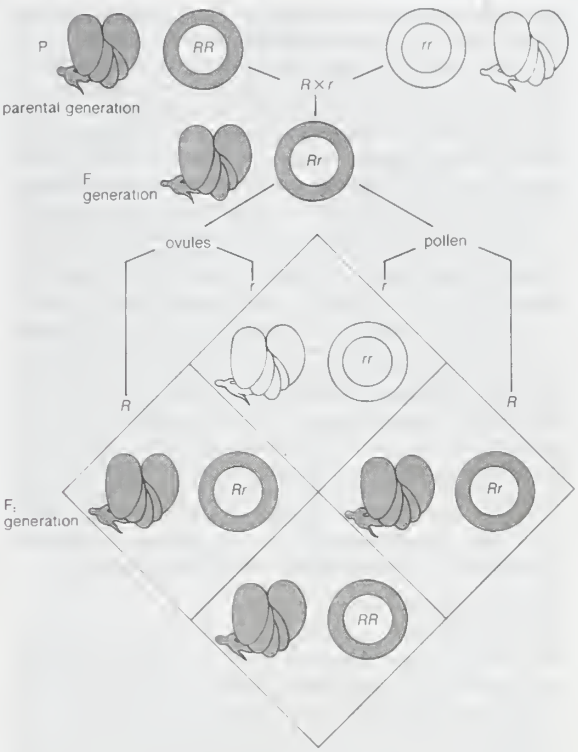


Figure 1: *Mendel's law of segregation.*
Cross of a purple-flowered and a white-flowered strain of peas. *R* stands for the gene for purple flowers and *r* for the gene for white flowers. Dark rings indicate the presence of a dominant gene for purple flowers.
From T. Dobzhansky, *Evolution, Genetics and Man* (1955), John Wiley & Sons, Inc.

First and second laws of Mendel

An example of one of Mendel's experiments will illustrate how the genes are transmitted and in what particular ratios. Let *R* stand for the gene for purple flowers and *r* for the gene for white flowers (dominant genes are conventionally symbolized by capital letters and recessive genes by small letters). Since each pea plant contains a gene endowment half of whose set is derived from the mother and half from the father, each plant has two genes for flower colour. If the two genes are alike, for instance, both having come from white-flowered parents (*rr*), the plant is termed a homozygote (Figure 1). The union of gametes with different genes give a hybrid plant termed a heterozygote (*Rr*). Since the gene *R*, for purple, is dominant over *r*, for white, the F_1 generation hybrids will show purple flowers. They are phenotypically purple, but their genotype contains both *R* and *r* genes, and these alternative (allelic or allelomorphic) genes do not blend or contaminate each other. Mendel inferred that when a heterozygote forms its sex cells, the allelic genes segregate and pass to different gametes. This is expressed in the first law of Mendel, the law of segregation of unit genes. Equal numbers of gametes, ovules, or pollen grains are formed that contain the genes *R* and *r*. Now, if the gametes unite at random, then the F_2 generation should contain about 1/4 white-flowered and 3/4 purple-flowered plants. The white-flowered plants, which must be recessive homozygotes, bear the genotype *rr*. About 1/3 of the plants exhibiting the dominant trait of purple flowers must be homozygotes, *RR*, and 2/3 heterozygotes, *Rr*. The prediction is tested by obtaining a third generation, F_3, from the purple-flowered plants; though phenotypically all purple-flowered, 2/3 of this group of plants reveal the presence of the recessive gene allele, *r*, in their genotype by producing about 1/4 white-flowered plants in the F_3 generation.

Mendel also crossbred varieties of peas that differed in two or more easily distinguishable traits. When a variety with yellow round seed was crossed to a green wrinkled seed variety (Figure 2), the F_1 generation hybrids produced yellow round seed. Evidently yellow (*A*) and round (*B*) are dominant traits, green (*a*) and wrinkled (*b*) are recessive. By allowing the F_1 plants (genotype *AaBb*) to self-pollinate, Mendel obtained an F_2 generation of 315 yellow round, 101 yellow wrinkled, 108 green round, and 32 green wrinkled seeds, a ratio approximately 9 : 3 : 3 : 1. The important point here is that the segregation of the colour (*A–a*) is independent of the segregation of the trait of seed surface (*B–b*). This is expected if the F_1 generation produces equal numbers of four kinds of gametes, carrying the four possible combinations of the parental genes: *AB*, *Ab*, *aB*, and *ab*. Random union of these gametes gives, then, the four phenotypes in a ratio 9 dominant–dominant : 3 recessive–dominant : 3 dominant–recessive : 1 recessive–recessive. Among these four phenotypic classes there must be nine different genotypes, a supposition that can be tested experimentally by raising a third hybrid generation. The predicted genotypes are actually found. Another test is by means of a backcross (or testcross)—the F_1 hybrid (phenotype yellow round seed; genotype *AaBb*) is crossed to a double recessive plant (phenotype green wrinkled seed; genotype *aabb*). If the hybrid gives four kinds of gametes in equal numbers and if all the gametes of the double recessive are alike (*ab*), the predicted progeny of the backcross are yellow round, yellow wrinkled, green round, and green wrinkled seed in a ratio 1 : 1 : 1 : 1. This prediction is realized in experiments. When the varieties crossed differ in three genes, the F_1 hybrid forms 2^3 or eight kinds of gametes (2^n = kinds of gametes, n being the number of genes). The second generation of hybrids, the F_2, has 27 (3^3) genotypically distinct kinds of individuals but only eight different phenotypes. From these results and others Mendel derived his second law, the law of recombination or independent assortment of genes.

UNIVERSALITY OF MENDEL'S LAWS

Although Mendel experimented with varieties of peas, his laws have been shown to apply to the inheritance of many kinds of characters in almost all organisms. In 1902 Mendelian inheritance was demonstrated in poultry and in mice. The following year, albinism became the first human trait shown to be a Mendelian recessive, with pigmented skin the corresponding dominant.

In 1902 and 1909 Sir Archibald Garrod initiated the analysis of inborn errors of metabolism in humans in terms of biochemical genetics. Alkaptonuria, inherited as a recessive, is characterized by excretion in the urine of large amounts of the substance called alkapton, or homogentisic acid, which renders the urine black on exposure to air. In normal (*i.e.*, nonalkaptonuric) persons the homogentisic acid is changed to acetoacetic acid, the reaction being facilitated by an enzyme, homogentisic acid oxidase. Garrod advanced the hypothesis that this enzyme is absent or inactive in homozygous carriers of the defective recessive alkaptonuria gene; hence the homogentisic acid accumulates and is excreted in the urine. Mendelian inheritance of numerous traits in humans has been studied since then (see below *Human genetics*).

From T Dobzhansky, *Evolution, Genetics and Man* (1955); John Wiley & Sons, Inc.

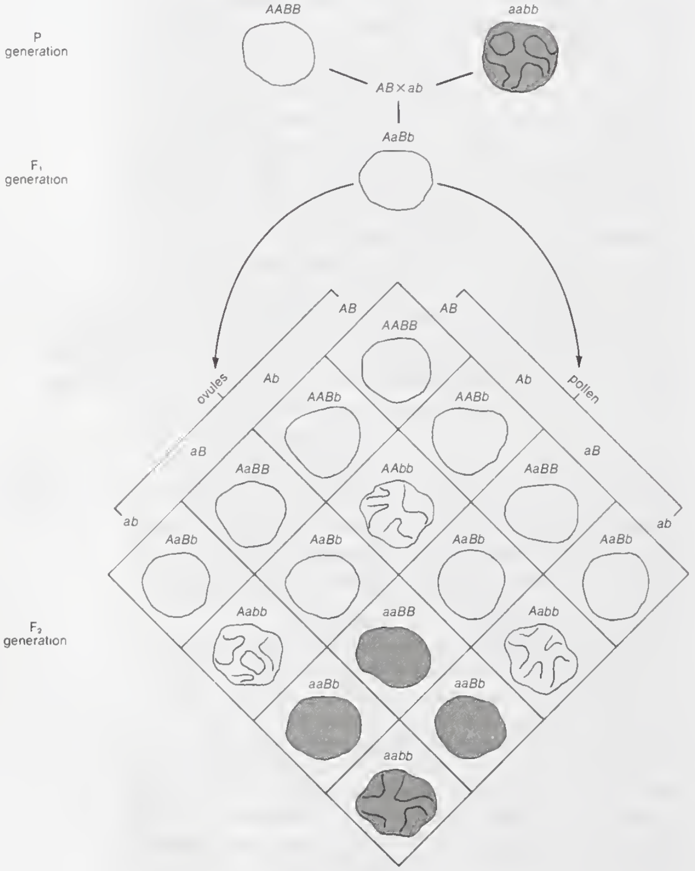


Figure 2: *Mendel's law of independent assortment.*
Cross of peas having yellow and smooth seeds with peas having green and wrinkled seeds. *A* stands for the gene for yellow and *a* for the gene for green; *B* stands for the gene for a smooth surface and *b* for the gene for a wrinkled surface.

In analyzing Mendelian inheritance, it should be borne in mind that an organism is not an aggregate of independent traits, each determined by one gene. A "trait" is really an abstraction, a term of convenience in description. One gene may affect many traits (a condition termed pleiotropic). The gene white in *Drosophila* flies is pleiotropic; it affects the colour of the eyes and of the testicular envelope in the males, the fecundity and the shape of the spermatheca in the females, and the longevity of both sexes. In humans many diseases caused by a single defective gene will have a variety of symptoms, all pleitropic manifestations of the gene.

ALLELIC INTERACTIONS

Dominance relationships. The operation of Mendelian inheritance is frequently more complex than in the case of the traits recorded by Mendel. In the first place, clear-cut dominance and recessiveness are by no means always found. When red- and white-flowered varieties of four-o'clock plants or snapdragons are crossed, for example, the F_1 hybrids have flowers of intermediate pink or rose colour, a situation that seems more explicable by the blending notion of inheritance than by Mendelian concepts. That the inheritance of flower colour is indeed due to Mendelian mechanisms becomes apparent when the F_1 hybrids are allowed to cross, yielding an F_2 generation of red-, pink-, and white-flowered plants in a ratio of 1 red : 2 pink : 1 white. Obviously the hereditary information for the production of red and white flowers had not been blended away in the first hybrid generation, as flowers of these colours were produced in the second generation of hybrids.

Ranges of dominance

The apparent blending in the F_1 generation is explained by the fact that the gene alleles that govern flower colour in four-o'clocks show an incomplete dominance relationship. Suppose, then, that a gene allele R_1 is responsible for red and R_2 for white flowers; the homozygotes R_1R_1 and R_2R_2 are red and white respectively, and the heterozygotes R_1R_2 have pink flowers. A similar pattern of lack of dominance is found in shorthorn cattle. In diverse organisms, dominance ranges from complete (a heterozygote indistinguishable from one of the homozygotes) through incomplete (heterozygotes exactly intermediate) to excessive or over-dominance (a heterozygote more extreme than either homozygote).

Another form of dominance is one in which the heterozygote displays the phenotypic characteristics of both alleles. This is called codominance; an example is seen in the MN blood group system of human beings. MN blood type is governed by two alleles, *M* and *N*. Individuals who are homozygous for the *M* allele have a surface molecule (called the M antigen) on their red blood cells. Similarly, those homozygous for the *N* allele have the N antigen on the red blood cells. Heterozygotes—those with both alleles—carry both antigens.

Multiple alleles. The traits discussed so far all have been governed by the interaction of two possible alleles. Many genes, however, are represented by multiple allelic forms within a population. (One individual, of course, can possess only two of these multiple alleles.) Human blood groups—in this case, the well-known ABO system—again provide an example. The gene that governs ABO blood types has three alleles: I^A, I^B, and I^O. I^A and I^B are codominant, but I^O is recessive. Because of the multiple alleles and their various dominance relationships, there are four phenotypic ABO blood types: type A (genotypes I^AI^A and I^AI^O), type B (genotypes I^BI^B and I^BI^O), type AB (genotype I^AI^B), and type O (genotype I^OI^O).

GENE INTERACTIONS

Many individual traits are affected by more than one gene. For example, the coat colour in many mammals is determined by numerous genes interacting to produce the result. The great variety of colour patterns in cats, dogs, and other domesticated animals is the result of different combinations of complexly interacting genes. The gradual unravelling of their modes of inheritance was one of the active fields of research in the early years of genetics.

Two or more genes may produce similar and cumulative effects on the same trait. In humans, the skin colour difference between so-called blacks and so-called whites is due to several (probably four or more) interacting pairs of genes, each of which increases or decreases the skin pigmentation by a relatively small amount.

Epistatic genes. Some genes mask the expression of other genes just as a fully dominant allele masks the expression of its recessive counterpart. The gene that masks the phenotypic effect of another gene is called the epistatic gene; the gene it subordinates is the hypostatic gene. The gene for albinism (lack of pigment) in humans is an epistatic gene. It is not part of the interacting skin-colour genes described above; rather, its dominant allele is necessary for the development of any skin pigment, and its recessive homozygous state results in the albino condition regardless of how many other pigment genes may be present. Albinism thus occurs in some individuals among people who belong to dark- or intermediate-pigmented groups, such as blacks and American Indians, as well as among whites.

Non-penetrance

The presence of epistatic genes explains much of the variability seen in the expression of such dominantly inherited human diseases as Marfan's syndrome and neurofibromatosis. Because of the effects of an epistatic gene, some individuals who inherit the dominant, disease-causing gene show only partial symptoms of the disease; some, in fact, may show no expression of the disease-causing gene, a condition referred to as nonpenetrance. The individual in whom such a nonpenetrant mutant gene exists will be phenotypically normal but still capable of passing the deleterious gene on to offspring, who may exhibit the full-blown disease.

Examples of epistasis abound in nonhuman organisms. In mice, as in humans, the gene for albinism has two variants: the allele for nonalbino and the allele for albino. The latter allele is unable to synthesize the pigment melanin. Mice, however, have another pair of alleles involved in melanin placement. These are the agouti allele, which produces dark melanization of the hair except for a yellow band at the tip, and the black allele, which produces melanization of the whole hair. If melanin cannot be formed (the situation in the mouse homozygous for the albino gene) neither agouti nor black can be expressed. Hence homozygosity for the albinism gene is epistatic to the agouti/black alleles and prevents their expression.

Complementation. The phenomenon of complementation is another form of interaction between nonallelic genes. For example, there are mutant genes that in the homozygous state produce profound deafness in humans. One would expect that the children of two persons suffering from such hereditary deafness would all be deaf. This is frequently not the case, because the parents' deafness is often caused by different genes. Since the mutant genes are not alleles, the child becomes heterozygous for the two genes and hears normally. In other words, the two mutant genes complement each other in the child. Complementation thus becomes a test for allelism. In the case of congenital deafness cited above, if all the children had been deaf, one could assume that the deafness in each of the parents was due to mutant genes that were alleles. This would be more likely to occur if the parents were genetically related (consanguineous).

Polygenic inheritance. The greatest difficulties of analysis and interpretation are presented by the inheritance of many quantitative or continuously varying traits. Inheritance of this kind produces variations in degree rather than in kind, in contrast to the inheritance of discontinuous traits resulting from single genes of major effect (see above). The yield of milk in different breeds of cattle, egg-laying capacity in poultry, and stature, shape of the head, blood pressure, and intelligence in humans range in continuous series from one extreme to the other and are significantly dependent on environmental conditions. Crosses of two varieties differing in such characters usually give F_1 hybrids intermediate between the parents. At first sight, this situation suggests a blending inheritance through "blood" rather than Mendelian inheritance; in fact, it was probably observations of this kind of inheritance that suggested the folk idea of "blood theory."

Mendelian segregation in polygenes

It has, however, been shown that these characters are polygenic—*i.e.*, determined by several or many genes, each taken separately producing only a slight effect on the phenotype, as small as or smaller than that caused by environmental influences on the same characters. That Mendelian segregation does take place with polygenes, as with the genes having major effects (sometimes called oligogenes), is shown by the variation among F_2 and further generation hybrids being usually much greater than that in the F_1 generation. By selecting among the segregating progenies the desired variants, for example, individuals or families with the greatest yield, the best size, or a desirable behaviour, it is possible to produce new breeds or varieties sometimes exceeding the parental forms. Hybridization and selection are consequently potent methods that can be used for improvement of agricultural plants and animals.

Polygenic inheritance also applies to many of the birth defects (congenital malformations) seen in humans. Although expression of the defect itself may be discontinuous (as in clubfoot, for example), susceptibility to the trait is continuously variable and follows the rules of polygenic inheritance. When a developmental threshold produced by a polygenically inherited susceptibility and a variety of environmental factors is exceeded, the birth defect results (see below *Genetics and human disease*).

Heredity and environment

PREFORMISM AND EPIGENESIS

The homunculus of early microscopists

A notion that was widespread among pioneer biologists in the 18th century was that the fetus, and hence the adult organism that develops from it, is preformed in the sex cells. Some early microscopists even imagined that they saw a tiny "homunculus," a diminutive human figure, encased in the human spermatozoon. The development of the individual from the sex cells appeared deceptively simple—it was merely increase in size and growth of what was already present in the sex cells. The antithesis of the early preformation theories were theories of epigenesis, which claimed that the sex cells were structureless jelly and contained nothing at all in the way of rudiments of future organisms. The naive early versions of preformation and epigenesis had to be given up when embryologists showed that the embryo develops by a series of complex but orderly and gradual transformations (see GROWTH AND DEVELOPMENT, BIOLOGICAL: *Animal development*). Darwin's "Provisional Hypothesis of Pangenesis" was distinctly preformistic; Weismann's theory of determinants in the germ plasm as well as the early ideas about the relations between genes and traits also tended toward preformism.

Heredity has been defined as a process that results in the progeny's resembling their parents. A further qualification of this definition states that what is inherited is a potential that expresses itself only after interacting with and being modified by environmental factors. In short, all phenotypic expressions have both hereditary and environmental components, the amount of each varying for different traits. Thus, a trait that is primarily hereditary (*e.g.*, skin colour in humans) may be modified by environmental influences (*e.g.*, suntanning). And conversely, a trait sensitive to environmental modifications (*e.g.*, weight in humans) is also genetically conditioned. Organic development is preformistic insofar as a fertilized egg cell contains a genotype that conditions the events that may occur and is epigenetic insofar as a given genotype allows a variety of possible outcomes. These considerations should dispel the reluctance felt by many people to accept the fact that mental as well as physiological and physical traits in humans are genetically conditioned. Genetic conditioning does not mean that heredity is the "dice of destiny." At least in principle, but not invariably in practice, the development of a trait may be manipulated by changes in the environment.

HERITABILITY

Although hereditary diseases and malformations are, unfortunately, by no means uncommon in the aggregate, no one of them occurs very frequently. The characteristics by which one person is distinguished from another, such as facial features, stature, shape of the head, skin, eye and hair colours, and voice, are not usually inherited in a clear-cut Mendelian manner, as are some hereditary malformations and diseases. This is not as strange as it may seem. The kinds of gene changes, or mutations, that produce morphological or physiological effects drastic enough to be clearly set apart from the more usual phenotypes are likely to cause diseases or malformations just because they are so drastic.

Variations caused by polygenes

The variations that occur among healthy persons are, as a general rule, caused by polygenes with individually small effects. The same is true of individual differences among members of various animal and plant species. Even brown-blue eye colour in humans, which in many families behaves as if caused by two forms of a single gene (brown being dominant and blue recessive), is often blurred by minor gene modifiers of the pigmentation. Some apparently blue-eyed persons actually carry the gene for the brown eye colour, but several additional modifier genes decrease the amount of brown pigment in the iris. This type of genetic process can influence susceptibility to many diseases (*e.g.*, diabetes) or birth defects (for example, cleft lip—with or without cleft palate).

From T Dobzhansky, *Heredity and the Nature of Man*, (1964); Harcourt Brace Jovanovich, Inc

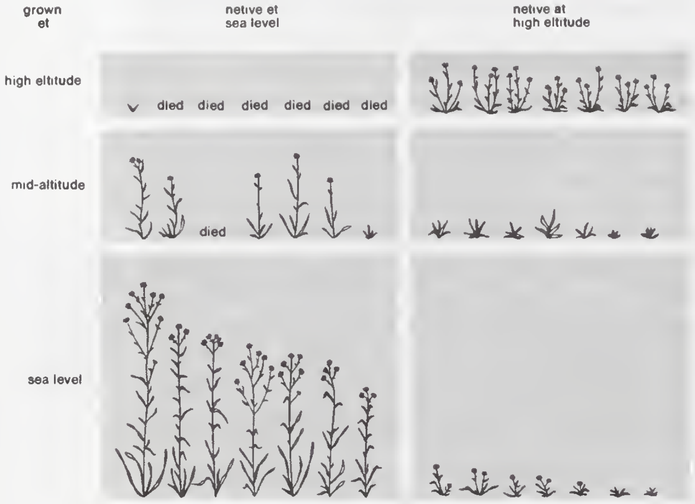


Figure 3: *Genotype–environment interaction.*
Yarrow plants native in different habitats divided and replanted at different elevations.

The question geneticists must often attempt to answer is how much of the observed diversity between persons, or between individuals of any species, is due to hereditary, or genotypic, variations and how much of it is due to environmental influences. Applied to human beings, this is sometimes referred to as the nature–nurture problem. (See *Human genetics* below, for a discussion of this aspect of the subject.) With animals or plants the problem is evidently more easily soluble than it is with people. Two complementary approaches are possible. First, individual organisms, or their progenies, are raised in environments as uniform as can be provided, with food, temperature, light, humidity, etc., carefully controlled. The differences that persist between such individuals or progenies probably reflect genotypic differences. Second, individuals with similar or identical genotypes are placed in different environments. The phenotypic differences then may be ascribed to environmental induction. Experiments combining both approaches have been carried out on several species of plants that grow naturally at different altitudes, from sea level to the alpine zone of the Sierra Nevada of California. Young yarrow plants (*Achillea*) were cut in three parts and the cuttings replanted in experimental gardens at sea level, at midaltitude (4,800 feet [1,460 metres]), and at a high altitude (10,000 feet [3,050 metres]) (Figure 3). The plants native at sea level grow best in their native habitat, grow less well at midaltitudes, and die at high altitudes. On the other hand, the alpine race survives and develops better at the high-altitude transplant station than it does at lower altitudes.

Table 2: Some Heritability Estimates

trait	correlation
Cattle	
Butterfat percentage	0.6
Milk yield	0.3
Pigs	
Body length	0.5
Weight at 180 days	0.3
Litter size	0.15
Poultry	
Egg weight	0.6
Annual egg production	0.3
Body weight	0.2
Viability	0.1
Drosophila melanogaster	
Abdominal bristle number	0.5
Body size	0.4
Ovary size	0.3
Egg production	0.2

Source: D.S. Falconer, *Introduction to Quantitative Genetics*, 1960.

With organisms that cannot survive being cut in pieces and placed in controlled environments, a partitioning of the observed variability into genetic and environmental components may be attempted by other methods. Suppose that in a certain population individuals vary in stature, weight, or some other trait. These characters can be measured in many pairs of parents and in their progenies raised under different environmental conditions. If the variation is due entirely to environment and not at all to heredity, then the expression of the character in the parents and in the offspring will show no correlation at all (heritability = zero). On the other hand, if the environment is unimportant and the character is uncomplicated by dominance, then the means of this character in the progenies will be the same as the means of the parents; with differences in the expression in females and in males taken into account, the heritability will equal unity. In reality, most heritabilities are found to lie between zero and one. Some examples of heritabilities of traits in different animals are given in Table 2.

Heritability estimates

It is important to understand clearly the meaning of heritability estimates. They show that, given the range of the environments in which the experimental animals lived, one could predict the average body sizes in the progenies of pigs better than one could predict the average numbers of piglets in a litter. The heritability is, however, not an inherent or unchangeable property of each character. If one could make the environments more uniform, the heritabilities would rise, and with more diversified environments they would decrease. Similarly, in populations that are more variable genetically the heritabilities increase, and in genetically uniform ones they decrease. In humans the situation is even more complex because the environments of the parents and of their children are in many ways interdependent. Suppose, for example, that one wishes to study the heritability of stature, weight, or susceptibility to tuberculosis. The stature, weight, and liability to contract tuberculosis depend to some extent on the quality of nutrition and generally on the economic well-being of the family. If no allowance is made for this fact, the heritability estimates arrived at may be spurious; such heritabilities have indeed been claimed for such things as administrative, legal, or military talents and for social eminence in general. It is evident that having socially eminent parents makes it easier for the children to achieve such eminence also; biological heredity may have little or nothing to do with this.

A general conclusion from the evidence now available may be stated as follows: diversity in almost any trait, physical, physiological, or behavioral, is due in part to genetic and in part to environmental variables. In any array of environments, individuals with more nearly similar genetic endowments are likely to show a greater average resemblance than the carriers of more diverse genetic endowments. It is, however, also true that in different environments the carriers of similar genetic endowments may grow, develop, and behave in different ways.

THE PHYSICAL BASIS OF HEREDITY

When Mendel formulated his laws of heredity, he postulated a particulate nature for the units of inheritance. What exactly these particles were he did not know. Today scientists understand not only the physical location of hereditary units (*i.e.,* the genes) but their molecular composition as well. The unravelling of the physical basis of heredity makes up one of the most fascinating chapters in the history of biology.

Chromosomes and genes

As has been discussed, each individual in a sexually reproducing species inherits two alleles for each gene, one from each parent. Furthermore, when such an individual forms sex cells, each of the resultant gametes receives one member of each allelic pair. The formation of gametes occurs through a process of cell division called meiosis; it is also known as reduction division, because the amount of hereditary material present in the gametes has been reduced by half. When gametes unite in fertilization, the double dose of hereditary material is restored, and a new individual is created. This individual, consisting at first of only one cell, grows via mitosis, a process of repeated cell divisions. Mitosis differs from meiosis in that each daughter cell receives a full copy of all the hereditary material found in the parent cell.

It is apparent that the genes must physically reside in cellular structures that meet two criteria. First, these structures must be replicated and passed on to each generation of daughter cells during mitosis. Second, they must be organized into homologous pairs, one member of which is parcelled out to each gamete formed during meiosis.

Discovery of chromosomes

As early as 1848, biologists had observed that cell nuclei resolve themselves into small, rodlike bodies during mitosis; later these structures were found to absorb certain dyes and so came to be called chromosomes (coloured bodies). During the early years of the 20th century, cellular studies using ordinary light microscopes clarified the behaviour of chromosomes during mitosis and meiosis, which led to the conclusion that chromosomes are the carriers of genes.

THE BEHAVIOUR OF CHROMOSOMES DURING CELL DIVISION

During mitosis. When the chromosomes condense during cell division, they have already undergone replication. Each chromosome thus consists of two identical replicas, called chromatids, joined at a point called the centromere (see Figure 4). During mitosis the sister chromatids separate, one going to each daughter cell. Chromosomes thus meet the first criterion for being the repository of genes: they are replicated and a full copy is passed to each daughter cell during mitosis.

During meiosis. It was the behaviour of chromosomes during meiosis, however, that provided the strongest evidence for their being the carriers of genes. In 1902 the American Walter S. Sutton reported on his observations of the action of chromosomes during sperm formation in grasshoppers. Sutton had observed that during meiosis, each chromosome (consisting of two chromatids) becomes paired with another, physically similar chromosome. These homologous chromosomes separate during meiosis, with one member of each pair going to a different cell. Assuming that one member of each homologous pair was of maternal origin and the other was paternally derived, here was an event that fulfilled the behaviour of genes postulated in Mendel's first law.

It is now known that the number of chromosomes within the nucleus is usually constant in all individuals of a given species—for example, 46 in humans; 40 in the house mouse; 8 in the vinegar, or fruit, fly (*Drosophila melanogaster*); 20 in corn (maize); 24 in the tomato; 48 in the potato. In sexually reproducing organisms, this number is called the diploid number of chromosomes, as it represents the double dose of chromosomes received from two parents. The nucleus of a gamete, however, contains half this number of chromosomes, or the haploid number. Thus a human gamete contains 23 chromosomes, while a *Drosophila* gamete contains four. Meiosis produces the haploid gametes.

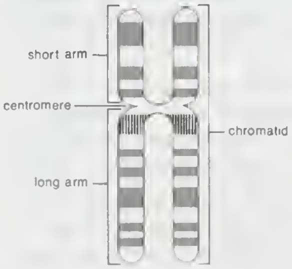


Figure 4: *Structure of the mitotic chromosome.*
Before mitosis, each chromosome has replicated so that it consists of two identical chromatids, which are joined at the centromere. The light and dark bands are produced by Giemsa-trypsin staining; the light bands are the primary sites of unique DNA, and the dark bands are primarily repetitive DNA. The chromatids separate by division through the centromere during mitosis; this division yields two identical chromosomes, one of which goes to each of the two daughter cells produced by mitosis.

The essential features of meiosis are shown in Figure 5. For the sake of simplicity, the diploid parent cell is shown to contain a single pair of homologous chromosomes, one member of which is represented black (from the father) and the other white (from the mother). At the leptotene stage the chromosomes appear as long, thin threads. At pachytene they pair, the corresponding portions of the two chromosomes lying side-by-side. The chromosomes then duplicate and contract into paired chromatids. At this stage the pair of chromosomes is known as a tetrad, as it consists of four chromatids. Also at this stage an extremely important event occurs: portions of the maternal and paternal chromosomes are exchanged. This exchange process, called crossing over, results in chromatids that include both paternal and maternal genes and consequently introduces new genetic combinations. The first meiotic, or reduction, division separates the chromosomal tetrads, with the paternal chromosome (whose chromatids now contain some maternal genes) going to one cell, and the maternal chromosome (containing some paternal genes) going to another cell. During the second meiotic division the chromatids separate. The original diploid cell has thus given rise to four haploid gametes (only two of which are shown in Figure 5). Not only has a reduction in chromosome number occurred, but the resulting single member of each homologous chromosome pair may be a new combination (through crossing over) of genes present in the original diploid cell.

Suppose that the white chromosome shown in Figure 5 carries the gene for albinism, and the black chromosome carries the gene for dark pigmentation. It is evident that

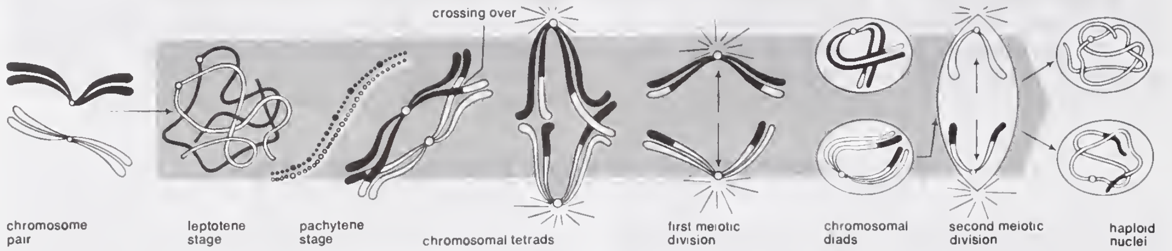


Figure 5: Behaviour of chromosomes at meiosis (see text).
From T. Dobzhansky, *Evolution, Genetics and Man*, (1955), John Wiley & Sons, Inc

the two gene alleles will undergo segregation at meiosis, and that one-half of the gametes formed will contain the albino gene and the other half the pigmentation gene. Following the scheme in Figure 1, random combination of the gametes with the albino and the pigmentation gene will give two kinds of homozygotes and one kind of heterozygote in a ratio 1 : 1 : 2. Mendel's law of segregation is, thus, the outcome of chromosome behaviour at meiosis. The same is true of the second law, that of the independent assortment.

Consider the inheritance of two pairs of genes, such as Mendel's factors for seed coloration and seed surface in peas; these genes are located on different pairs of chromosomes. Since maternal and paternal members of different chromosome pairs are assorted independently, so are the genes they contain. This explains, in part, the genetic variety seen among the progeny of the same pair of parents. As stated above, humans have 46 chromosomes in the body cells and in the cells (oogonia and spermatogonia) from which the sex cells arise. At meiosis these 46 chromosomes form 23 pairs, one of the chromosomes of each pair being of maternal and the other of paternal origin. Independent assortment is, then, capable of producing 2^{23}, or 8,388,608, kinds of sex cells with different combinations of the grandmaternal and grandpaternal chromosomes. Since each parent has the potentiality of producing 2^{23} kinds of sex cells, the total number of possible combinations of the grandparental chromosomes is $2^{23} \times 2^{23} = 2^{46}$. The population of the world is now more than 4,000,000,000 persons, or approximately 2^{32} persons. It is, therefore, certain that only a tiny fraction of the potentially possible chromosome and gene combinations can ever be realized. Yet even 2^{46} is an underestimate of the variety potentially possible. The grandmaternal and grandpaternal members of the chromosome pairs are not indivisible units. Each chromosome carries many genes, and the chromosome pairs exchange segments at meiosis through the process of crossing over. This is evidence that the genes rather than the chromosomes are the units of Mendelian segregation.

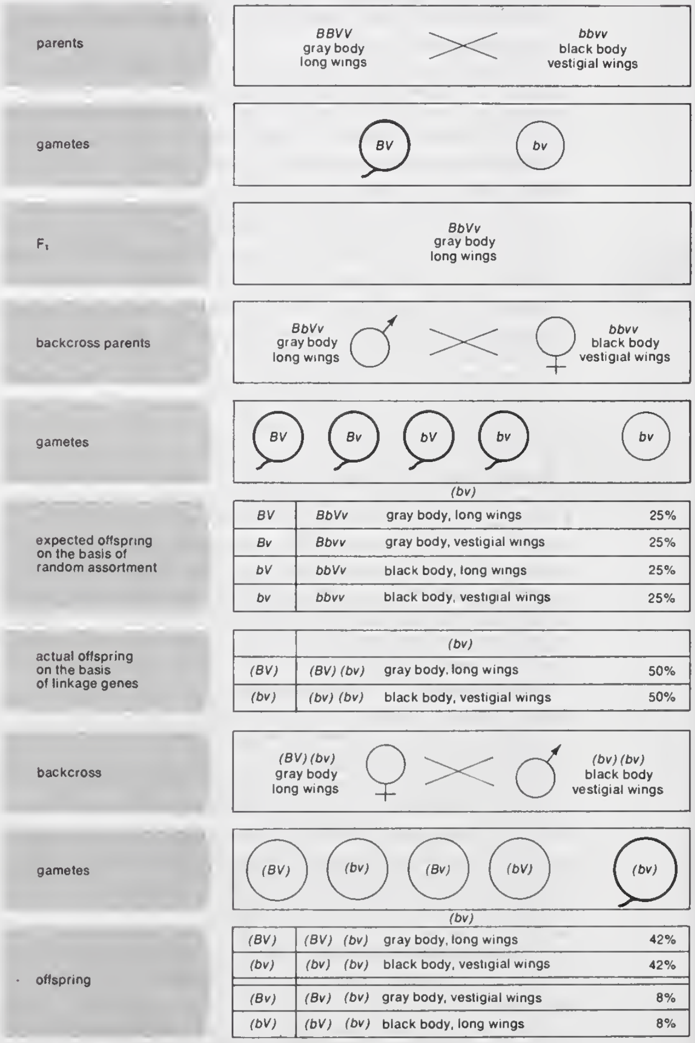
LINKAGE OF TRAITS

Simple linkage. As pointed out above, the random assortment of the maternal and paternal chromosomes at meiosis is the physical basis of the independent assortment of genes and of the traits they control. This is the basis of the second law of Mendel. The number of the genes in a sex cell is, however, much greater than that of the chromosomes. When two or more genes are borne on the same chromosome these genes may not be assorted independently; such genes are said to be linked. When a *Drosophila* fly homozygous for a normal gray body and long wings is crossed with one having a black body and vestigial wings, the F_1 consists of hybrid gray, long-winged flies (Figure 6). Gray body (B) is evidently dominant over black body (b), and long wing (V) is dominant over vestigial wing (v). Now consider a backcross of the heterozygous F_1 males to double recessive black-vestigial females ($bbvv$). Independent assortment would be expected to give in the progeny of the backcross the following: 1 gray-long : 1 gray-vestigial : 1 black-long : 1 black-vestigial. In reality only gray-long and black-vestigial flies are produced, in approximately equal numbers; the genes remain linked in the same combinations in which they were found in the parents. The backcross of the heterozygous F_1 females to double recessive males gives a somewhat different result: 42 percent each of gray-long and black-vestigial flies and about 8 percent each of black-long and gray-vestigial classes. In sum, 84 percent of the progeny have the parental combinations of traits, and 16 percent have the traits recombined. The interpretation of these results given in 1911 by the U.S. geneticist Thomas Hunt Morgan laid the foundations of the theory of linear arrangement of genes in the chromosomes.

Linked genes in *Drosophila*

Traits that exhibit linkage in experimental crosses (such as black body and vestigial wings) are determined by genes located in the same chromosome. As more and more genes became known in *Drosophila melanogaster,* they fell neatly into four linkage groups corresponding to the four pairs of the chromosomes this species possesses. One linkage group consists of sex-linked genes, located in the X chromosome (see below); of the three remaining linkage groups, two have many more genes than the remaining one; this corresponds to the presence of two pairs of large

Reprinted by permission of the publisher, from L.H. Snyder and P.R. David, *The Principles of Heredity* (Lexington, Mass.: D.C. Heath and Company, 1957)



BV	BbVv	gray body, long wings	25%
Bv	Bbvv	gray body, vestigial wings	25%
bV	bbVv	black body, long wings	25%
bv	bbvv	black body, vestigial wings	25%

	(bv)		
(BV)	(BV) (bv)	gray body, long wings	50%
(bv)	(bv) (bv)	black body, vestigial wings	50%

(BV)	(BV) (bv)	gray body, long wings	42%
(bv)	(bv) (bv)	black body, vestigial wings	42%
(Bv)	(Bv) (bv)	gray body, vestigial wings	8%
(bV)	(bV) (bv)	black body, long wings	8%

Figure 6: Linkage of genes as illustrated by body colour and wing length in *Drosophila* flies (see text).

chromosomes and one pair of tiny dotlike chromosomes. The numbers of linkage groups in other organisms are equal to or smaller than the numbers of the chromosomes in the sex cells; *e.g.,* 10 linkage groups and 10 chromosomes in corn, 19 linkage groups and 20 chromosomes in the house mouse, 23 linkage groups and 23 chromosomes in human beings.

As seen above, the linkage of the genes black and vestigial in *Drosophila* is complete in heterozygous males, while in the progeny of females there appear about 17 percent of recombination classes. With very rare exceptions, the linkage of all genes belonging to the same linkage group is complete in *Drosophila* males, while in the females different pairs of genes exhibit all degrees of linkage from complete (no recombination) to 50 percent (random assortment). Morgan's inference was that the degree of linkage depends on physical distance between the genes in the chromosome: the closer the genes the tighter the linkage, and vice versa. Furthermore, Morgan perceived that the chiasmata, crosses that occur in meiotic chromosomes, indicate the mechanism underlying the phenomena of linkage and crossing over. As shown schematically in Figure 5, the maternal and paternal chromosomes (represented black and white) cross over and exchange segments, so that a chromosome emerging from the process of meiosis may consist of some maternal (grandmaternal) and some paternal (grandpaternal) sections. If the probability of crossing-over taking place is uniform along the length of a chromosome (which was later shown to be not quite true), then genes close together will be recombined less frequently than those far apart.

Chromosome maps

This realization opened an opportunity to map the arrangement of the genes and the estimated distances between them in the chromosome by studying the frequencies of recombination of various traits in the progenies of hybrids. In other words, the linkage maps of the chromosomes are really summaries of many statistical observations on the outcomes of hybridization experiments. In principle at least, such maps could be prepared even if the chromosomes, not to speak of the chiasmata at meiosis, were unknown. But an interesting and relevant fact is that in *Drosophila* males the linkage of the genes in the same chromosome is complete, and observations under the microscope show that no chiasmata are formed in the chromosomes at meiosis. In most organisms, including humans, chiasmata are seen in the meiotic chromosomes in both sexes, and observations on hybrid progenies show that recombination of linked genes occurs also in both sexes.

The most detailed chromosome maps have been constructed for Morgan's classical material—*Drosophila melanogaster.* Less detailed chromosome maps exist for some other species of *Drosophila* flies, for corn, the house mouse, the bread mold *Neurospora crassa,* and for some bacteria and bacteriophages (viruses that infect bacteria). Until quite late in the 20th century, the mapping of human chromosomes presented a particularly difficult problem: experimental crosses could not be arranged in humans and only a few linkages could be determined by analysis of unique family histories. However, the development of somatic cell genetics (see below) provided new understanding of human genetic processes and new methods of research. Using the techniques of somatic cell genetics, hundreds of genes have been mapped to the human chromosomes and many linkages established.

Sex linkage. The male of many animals has one chromosome pair, the sex chromosomes, consisting of unequal members called X and Y. At meiosis the X and Y chromosomes first pair, then disjoin and pass to different cells. One-half of the gametes (spermatozoa) formed contain the X and the other half the Y chromosome. The female has two X chromosomes, and all egg cells normally carry a single X. The eggs fertilized by X-bearing spermatozoa give females (XX), and those fertilized by Y-bearing spermatozoa give males (XY).

The genes located in the X chromosomes exhibit what is known as sex-linkage or crisscross inheritance. This is due to a crucial difference between the paired sex chromosomes and the other pairs of chromosomes (called autosomes). The members of the autosome pairs are truly homologous; that is, each member of a pair contains a full complement of the same genes (albeit, perhaps, in different allelic forms). The sex chromosomes, on the other hand, do not constitute a homologous pair, as the X chromosome is much larger and carries far more genes than does the Y. Consequently, many recessive alleles carried on the X chromosome of a male will be expressed just as if they were dominant, for the Y chromosome carries no genes to counteract them. The classic case of sex-linked inheritance, described by Morgan in 1910, is that of the white eyes in *Drosophila* (Figure 7). White-eyed females crossed to males with the normal red eye colour produce red-eyed daughters and white-eyed sons in the F_1 generation and equal numbers of white-eyed and red-eyed females and males in the F_2 generation. The cross of red-eyed females to white-eyed males gives a different result: both sexes are red eyed in F_1, the females in the F_2 generation are red eyed, half of the males are red eyed, and the other half white eyed. As interpreted by Morgan, the gene that determines the red or white eyes is borne on the X chromosome, and the allele for red eye is dominant over that for white eye. Since a male receives its single X chromosome from his mother, all sons of white-eyed females also have white eyes. A female inherits one X chromosome from her mother and the other X from her father. Red-eyed females may have genes for red eyes in both of their X chromosomes (homozygotes) or may have one X with the gene for red and the other for white (heterozygotes). In the progeny of heterozygous females one half of the sons will receive the X chromosome with the gene for white and will have white eyes, and the other half will receive the X with the gene for red eyes. The daughters of the heterozygous females crossed with white-eyed males will have either two X chromosomes with the gene for white and hence white eyes or will have one X with white and the other X with the gene for red eyes and will be red-eyed heterozygotes.

Sex-linked genes

In humans, the red-green colour blindness and hemophilia are among many traits showing sex-linked inheritance and consequently are due to genes borne in the X chromosome. Sex-linked human diseases are discussed in detail below (see *Diseases caused by single-factor [Mendelian] inheritance*).

From *Life: An Introduction to Biology* by Simpson, Pittendrigh and Tiffany, Copyright © 1957 by Harcourt Brace Jovanovich Inc. and reproduced with their permission

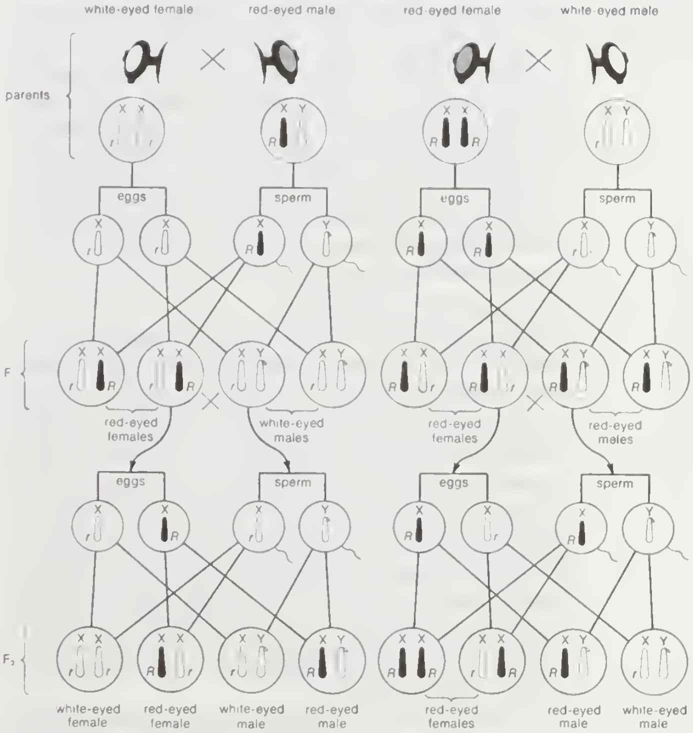


Figure 7: Sex-linked inheritance of white eyes in *Drosophila* flies (see text).

In some animals—birds, butterflies and moths, some fish, and at least some amphibians and reptiles—the chromosomal mechanism of sex determination is a mirror image of that described above. The male has two X chromosomes and the female an X and Y chromosome. Here the spermatozoa all have an X chromosome; the eggs are of two kinds, some with X and others with Y chromosomes, usually in equal numbers. The sex of the offspring is then determined by the egg rather than by the spermatozoon. Sex-linked inheritance is altered correspondingly. A male homozygous for a sex-linked recessive trait, crossed to a female with the dominant one, gives in the F_1 generation daughters with the recessive trait and heterozygous sons with the corresponding dominant trait. The F_2 generation has recessive and dominant females and males in equal numbers. A male with a dominant trait crossed to a female with a recessive trait gives uniformly dominant F_1 and a segregation in a ratio of 2 dominant males : 1 dominant female : 1 recessive female.

Observations on pedigrees or experimental crosses show that certain traits exhibit sex-linked inheritance; the behaviour of the X chromosomes at meiosis is such that the genes they carry may be expected to exhibit sex-linkage. This evidence still failed to convince some skeptics that the genes for the sex-linked traits were in fact borne in certain chromosomes seen under the microscope. An elegant experimental proof was furnished in 1916 by the U.S. geneticist Calvin Blackman Bridges. As stated above, white-eyed *Drosophila* females crossed to red-eyed males usually produce red-eyed female and white-eyed male progeny. Among thousands of such "regular" offspring there are occasionally found exceptional white-eyed females and red-eyed males. Bridges constructed the following working hypothesis. Suppose that during meiosis in the female, gametogenesis occasionally goes wrong, and the two X chromosomes fail to disjoin (Figure 8). Exceptional eggs will then be produced carrying two X chromosomes and eggs carrying none. An egg with two X chromosomes coming from a white-eyed female fertilized by a spermatozoon with a Y chromosome will give an exceptional white-eyed female. An egg with no X chromosome fertilized by a spermatozoon with an X chromosome derived from a red-eyed father will yield an exceptional red-eyed male. This hypothesis can be rigorously tested. The exceptional white-eyed females should have not only the two X chromosomes but also a Y chromosome, which normal females do not have. The exceptional males should, on the other hand, lack a Y chromosome, which normal males do have. Both predictions were verified by examination under a microscope of the chromosomes of exceptional females and males. The hypothesis also predicts that exceptional eggs with two X chromosomes fertilized by X-bearing spermatozoa must give individuals with three X chromosomes; such individuals were later identified by Bridges as poorly viable "superfemales." Exceptional eggs with no Xs, fertilized by Y-bearing spermatozoa, will give zygotes without X chromosomes; such zygotes die in early stages of development.

Verification of sex-linked inheritance

CHROMOSOMAL ABERRATIONS

Two general types of chromosomal abnormalities occur: numerical and structural. Numerical aberrations result from nondisjunction; that is, from the failure of a pair of homologous chromosomes or a pair of sister chromatids to separate during cell division. As described above, when nondisjunction occurs during meiosis two types of germ cells will be formed, those with an extra chromosome and those with a missing chromosome. If one of the former combines with a normal germ cell, the new fertilized egg and all the cells of the individual it produces will have an extra chromosome (trisomy); if one of the latter combines with a normal germ cell, the fertilized egg will lack a chromosome (monosomy). If nondisjunction occurs after fertilization, the resulting individual will be a mosaic and will have two or more populations of cells differing in chromosomal number.

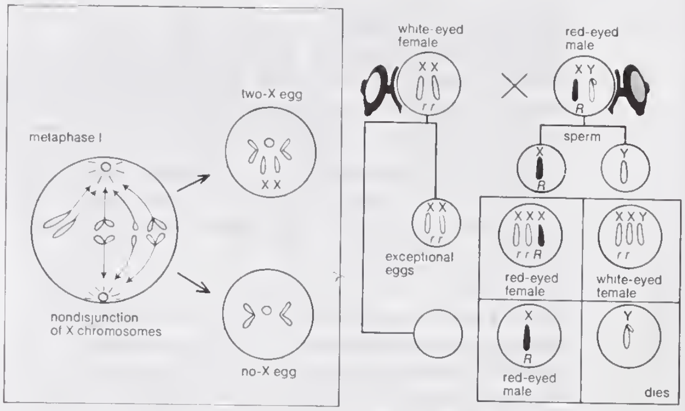


Figure 8: Nondisjunction of the X chromosomes in *Drosophila*, an explanation for the appearance of unexpected types of offspring. (Left) Abnormal meiosis in the female is responsible for (right) the exceptional eggs.

From *Life: An Introduction to Biology* by Simpson, Pittendrigh and Tiffany, Copyright © 1957 by Harcourt Brace Jovanovich, Inc., and reproduced with their permission

Structural aberrations result from chromosome breakages. Chromosomes may break spontaneously, or they may be broken by such environmental agents as radiation, viruses, and toxic chemicals. If a chromosomal segment breaks off and is not rejoined, it may be lost entirely in the gametes or somatic cells that derive, respectively, from meiosis or mitosis. Such a loss is called a deletion. In other instances, the broken-off segment may rejoin its chromosome but with its position inverted 180°; such inversions can alter the sequence of genetic information along the chromosome. In other cases, the segment may become translocated; that is, it may become attached to a different chromosome. When such a rearrangement occurs between two nonhomologous chromosomes without net loss or gain of chromosomal material, it is called a balanced, or reciprocal, translocation, and the individual is not phenotypically affected. If, however, the translocation results in the deletion or duplication of chromosomal material in gametes or somatic cells, the effects may be severe. This is especially true in the event of gametes that carry autosomal translocations; such chromosomal aberrations often produce lethal phenotypic effects.

(T.D./Ar.R.)

Molecular genetics

The data accumulated by the geneticists of the early 20th century provided compelling evidence that chromosomes are the carriers of the genes. But the nature of the genes themselves remained a mystery, as did the mechanism by which they exert their influence. Molecular genetics—the study of the molecular structure of the genes and the methods by which genes control the activities of the cell—provided the answers to these fundamental questions.

Much of the information in molecular genetics has come from the study of microorganisms, particularly the bacterium *Escherichia coli* (a common inhabitant of the human intestine) and its interactions with various bacteriophages. Bacteria have many features that make them especially useful in genetics research. For example, they have an extremely short life cycle, so that many generations can be raised in a brief period of time. Equally important, bacteria have only one basic function—to reproduce. Consequently, their genome is relatively limited. Furthermore, unlike most higher organisms, bacteria are not diploid, so their genome does not include two alleles of each gene. This makes it easy to identify a bacterium that carries a mutant gene, as the effects of the mutation cannot be masked by a normal allele. Although they are not diploid, bacteria can and do occasionally exchange genetic information through a variety of processes. This genetic exchange feature has been important in certain lines of molecular genetics research. Viruses also have advantages in genetics studies. Although they can reproduce only in a living cell, they have the simplest form of genetic material and evidence both genetically controlled properties and the ability to mutate.

Because of the relative simplicity of gene action in microorganisms, their study profoundly influenced early understanding of molecular genetics. Studies of the genetics of microorganisms involves the production of specific gene mutations and the examination of their biochemical effects. These studies have permitted the delineation of the metabolic pathways that produced the mutation in the experimental microorganism, as well as the isolation of the large molecules that contain the genetic information.

Although there are virtues to bacteria as experimental subjects in genetics research, it should be pointed out that bacteria differ from higher organisms in some rather fundamental ways. In fact, bacteria (along with the cyanophytes, or blue-green algae) are sufficiently distinct as to constitute their own kingdom, the Monera. Monerans, unlike protists, plants, and animals, are procaryotic. This means that their cells lack a true, membrane-enclosed nucleus, the cellular structure that contains the chromosomes in all other organisms (which are known as eucaryotes). Perhaps more important in a discussion of genetics, the bacterial chromosome differs in composition from the chromosomes of eucaryotes, so much so that some authorities prefer to avoid the term chromosome in describing the genetic material of bacteria. In eucaryotes, the chromosomes consist primarily of deoxyribonucleic acid (DNA) and a variety of proteins. Bacterial chromosomes have little protein, which proved to be an important clue in determining the chemical nature of the hereditary substance. Finally, all the progeny of a bacterium are identical, whereas the cell progeny of the fertilized egg of a complex, multicellular organism gives rise to many different tissues and organs whose component cells display specific patterns of different gene activities. This latter process is called differentiation.

HEREDITY AND NUCLEIC ACIDS

One of the most impressive and spectacular advances of biology in the 20th century was the discovery of the nature of the genetic material. The way information is encoded in the genes has been clarified and much has been learned about the mechanisms that translate this information into the developmental processes of the organism.

In 1869 a substance containing nitrogen and phosphorus was extracted from cell nuclei. It was originally called nuclein, but is now known as DNA. DNA is the chemical component of the chromosomes that is chiefly responsible for their staining properties in microscopic preparations. As stated above, the chromosomes of eucaryotes contain a variety of proteins in addition to DNA. The question naturally arose whether the nucleic acids or the proteins, or both together, are the carriers of the genetic information, which makes the genes of the same organism and of different organisms specifically different. Until the early 1950s most biologists were inclined to believe that the proteins were the chief carriers of heredity. Nucleic acids contain only four different unitary building blocks, but proteins are made up of 20 different amino acids. Proteins therefore appeared to have a greater diversity of structure, and the diversity of the genes seemed at first likely to rest on the diversity of the proteins.

The evidence that DNA acts as the carrier of the genetic information was first firmly demonstrated by exquisitely simple microbiological studies. One of these seminal studies was performed by the U.S. geneticist Oswald T. Avery and his coworkers in 1944. The background of Avery's study goes back to 1928, to research conducted by Fred Griffith of England, a bacteriologist who was studying the virulence of pneumococci, the bacteria that cause bacterial pneumonia. Griffith knew that the virulence of pneumococci—that is, their ability to cause infection—depends on the presence of an envelope composed of polysaccharides (sugar subunits) surrounding the bacterial cells. When grown on laboratory culture mediums, virulent pneumococci produced large colonies with a smooth, glistening surface.Bacteria from such cultures caused infection in mice. After many transfers to fresh laboratory mediums, however, some of the bacteria lost their polysaccharide envelopes and their ability to infect mice; correlated with these changes was the change to small colonies with rough outlines. The smooth and the rough variants were designated S and R, respectively. Griffith found that mice inoculated with either the living R pneumococci or with heat-killed S pneumococci remained free of infection, but mice inoculated with a mixture of living R and heat-killed S bacteria became infected. He further discovered that living S pneumococcal cultures could be obtained from such animals, proof that the virulent S cells were reconstituted from the mixture inoculated. Some material derived from the dead S bacteria had induced the transformation of R into S strains. Avery and his coworkers showed that the "transforming factor" which conferred the characteristic of virulence upon nonvirulent pneumococci consisted of DNA, which had been transferred from dead virulent pneumococci into living nonvirulent pneumococci. That this newly acquired characteristic was due to a specific genetic activity was demonstrated by its persistence in all of the offspring of the transformed bacteria.

Effect of the DNA of dead bacteria on living bacteria

The DNA of the dead cells evidently accomplishes the transformation of the living ones by penetrating the wall of the living cell. Once a section of the transforming DNA strand is inside the recipient cell, there apparently occurs a pairing between homologous regions of the bacterial chromosome and the transforming DNA. There must follow breakage and subsequent reunion of the bacterial chromosome and the transforming DNA. Thus a portion of the transforming DNA becomes integrated into the bacterial chromosome. If this model is valid, one would expect that genes located near each other on the transforming DNA would appear together more often in a transformed cell than will genes relatively far apart in the transforming DNA. This expectation has been fulfilled, and the principle has been utilized as a means of mapping the donor-cell chromosome. In further research the principles involved in transformation have been confirmed in a more efficient process involving mammalian cells in culture. By means of a process called transfection, defined pieces of DNA enter the cell nucleus and are incorporated into the DNA, thus replacing a particular genetic deficiency formerly exhibited by the cell.

Transfection

In the early 1950s Alfred D. Hershey and Martha Chase obtained evidence confirming that DNA serves as the physical basis of heredity. In their experiment, Hershey and Chase used a bacteriophage that infects *Escherichia coli*, a colon bacteria. This bacteriophage (or simply phage) is an ultramicroscopic tadpole-shaped particle, with a hexagonal head, a cylindrical tail, and an end plate with six tail fibres (see VIRUSES). The entire outer surface consists of protein, but within the interior space of the head there is DNA. When a phage infects *E. coli*, it injects its own genetic material into the bacterial cell. The phage genes then subvert the metabolic machinery of the bacterium, causing the host cell to make phage DNA and phage protein. When a new generation of phage particle is ready inside the host, they destroy (lyse) the bacterium. This lysis releases the new phage particles into the medium, where they can attack other bacterial cells.

Hershey and Chase prepared two populations of phage particles. In one population the phage protein was labelled with a radioactive isotope; in the other the phage DNA was radioactively labelled. After allowing both populations to attack *E. coli* cells, the experimenters analyzed the exterior and the interior of the infected cells for the presence of radioactive material. They discovered that the phage protein had remained outside of the host cell, while the phage DNA had been injected into the bacterium. This ingenious research demonstrated that the genetic material of the phage consists of DNA rather than protein.

The evidence is now overwhelming that the basic material constituting the gene is fundamentally the same in all organisms: it consists of chainlike molecules of nucleic acids—DNA in most organisms and RNA (ribonucleic acid, a close chemical relative of DNA) in certain viruses. As will be discussed later, the gene no longer stands for a discrete unit of heredity of definite and invariable length but is thought of as an operational entity whose properties are more fluid and depend upon the mode of measurement.

STRUCTURE OF NUCLEIC ACIDS

The remarkable properties of the nucleic acids, which qualify these substances to serve as the carriers of genetic information, have claimed the attention of many investigators. The groundwork was laid by pioneer biochemists who found that nucleic acids are long chainlike molecules, the backbones of which consist of repeated sequences of phosphate and sugar linkages—ribose sugar in RNA and deoxyribose sugars in DNA. Attached to the sugar links in the backbone are two kinds of nitrogenous bases: purines and pyrimidines. The purines are adenine (A) and guanine (G) in both DNA and RNA; the pyrimidines are cytosine (C) and thymine (T) in DNA and cytosine (C) and uracil (U) in RNA. A single purine or a pyrimidine is attached to each sugar, and the entire phosphate–sugar–base subunit is called a nucleotide. The nucleic acids extracted from different species of animals and plants have different proportions of the four nucleotides. Some are relatively richer in adenine and thymine, while others have more guanine and cytosine; however, the ratios of A to T, and also of G to C, are equal.

The Watson–Crick model of DNA

With the general acceptance of DNA as the chemical basis of heredity in the early 1950s, many microbiologists turned their attention to determining the molecular structure of this substance. In 1953 James Watson and Francis Crick proposed their now-famous model, which shows DNA as composed of two spirally wound (*i.e.*, helical) chains, in which the A's of one chain are linked by hydrogen bonds to the T's of the other, and the G's in one chain are linked to the C's of the other. The model looks something like a twisted ladder: the sides of the ladder are composed of the sugar and phosphate groups, while the rungs are made up of the paired nitrogenous bases. Watson and Crick based their model largely on X-ray crystallographic studies of DNA, which had been performed by Maurice Wilkins and Rosalind Franklin. This model fulfills the basic requirements—it makes it possible to envisage how genes replicate their precise structures when their copies are synthesized. It also makes it possible to explain how a gene can carry genetic information written in some chemical code. And, finally, it helps to envisage how mutational changes in the genes are produced.

DNA REPLICATION

The Watson–Crick model of the genetic material permits an explanation of the mechanism of precise replication of genes (Figure 9). The paired complementary strands of the DNA molecule may separate as a result of a breakage of the hydrogen bonds between the paired nitrogenous bases. If free nucleotides (base + sugar + phosphate) are present in the medium surrounding the gene, they might pair with the complementary bases on the single strands of DNA. An enzyme, DNA polymerase, functions to form the phosphate bonds between the sugars in the DNA backbone. It has been used in the synthesis of DNA in vitro, in cell-free systems. The enzyme is extracted from rapidly dividing cells of *E. coli*. A supply of the four nucleotides, A, T, G, and C, is provided, as well as a source of energy, adenosine triphosphate (ATP). To start DNA synthesis another key component is added—a trace of DNA to serve as a primer, or template. The kind of DNA that is synthesized depends on the primer. Even though the enzyme came from *E. coli*, if the primer is DNA of some quite different organism, such as cattle, the DNA that is synthesized is not *E. coli* but cattle DNA.

From *Molecular Biology of Bacterial Viruses* by Gunther S. Stent W.H. Freeman and Company, copyright © 1963

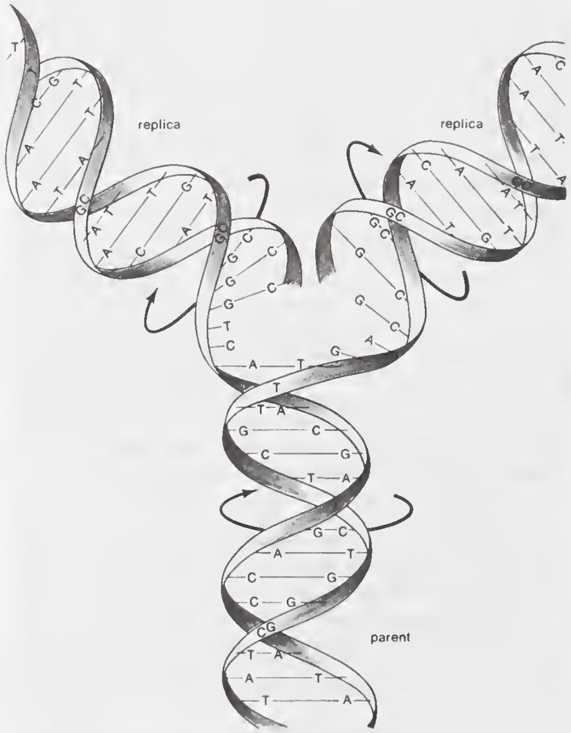


Figure 9: A possible method for the replication of DNA according to Watson and Crick. The arrows indicate the direction of rotation.

The Watson–Crick model of the structure of DNA suggested several different ways that DNA might self-replicate. The experiments of Matthew Meselson and Franklin Stahl on *E. coli* in 1958 suggested that DNA replicates semiconservatively to form two helices, each with one old and one new strand, the old strand acting as a template for the formation of a new strand. J. Herbert Taylor, using cells of higher organisms, confirmed the validity of this interpretation. Synthesis of new DNA occurs while the old chromosome is in the process of uncoiling (see Figure 9).

The DNA in one human cell is approximately two metres long when stretched out. It has been estimated that if all the DNA in a human were stretched out, it would extend from the Earth to the Sun and back again. For the large amount of DNA in one cell to fit, it obviously must be carefully and tightly packaged. About 140 base pairs of the DNA helix wind around a cluster of chromosome proteins (histones) to form a nucleosome, a structure similar to a bead on a string. Between the nucleosome beads is a string (linker region) of 20 to 100 DNA base pairs associated with another histone protein. This structure is flexible enough to permit the coiling and folding necessary to pack the DNA into the cell nucleus in a way that makes it readily available when it becomes genetically active.

DNA AS AN INFORMATION CARRIER: TRANSCRIPTION AND TRANSLATION OF THE GENETIC CODE

As has been stated, the Watson–Crick model provides an explanation of how a gene can carry hereditary information in the form of a chemical code. This section will describe the genetic code and explain how it governs the biochemical processes of the cell.

Before turning to the language of the code, it is necessary to explain what it is that the code specifies. It is now known that genes encode instructions for the production of proteins, which are largely responsible for the structure and function of the organism. Proteins are large, complex molecules consisting of one or more polypeptide chains that, in turn, are composed of amino acids linked together by peptide bonds. Proteins play many roles in organisms. Some proteins make up structural components of the organism; an example is the protein collagen in vertebrate animals. Others perform particular functions; for example, the protein hemoglobin transports oxygen in the blood of mammals, and the proteins of the immune system (immunoglobulins) protect against diseases in many members of the animal kingdom. Still other proteins regulate the rate of specific biochemical reactions in cells. This latter class of proteins, called enzymes, functions as biological catalysts. Enzymes permit chemical reactions to occur with extreme rapidity at temperatures normal to living cells. Without these proteins, the molecular interactions would require much longer periods of time and much higher temperatures, and they would lose their specificity. It is certainly no exaggeration to say that life depends on enzymes.

DNA and RNA: transcription. Among eucaryotes, DNA never leaves the cell nucleus, despite the fact that protein synthesis takes place on the ribosomes, structures that lie in the cytoplasm (*i.e.*, in the portion of the cell outside of the nucleus). Even among procaryotes, which have no membrane-enclosed nucleus, the DNA does not directly carry its instructions to the ribosomes. In both kinds of organisms, this function is performed by a type of RNA that copies the DNA message and carries it to the site of protein synthesis. Aptly enough, this RNA is called messenger RNA, or mRNA for short. The copying of the DNA instructions into messenger RNA is called the transcription function of DNA, to distinguish it from the replication function discussed above.

The sequence of the genetic letters, A (adenine), T (thymine), C (cytosine), and G (guanine), in the DNA is first transcribed into the corresponding sequence of the letters A, U (uracil), C, and G in the messenger RNA. This occurs through the action of the enzyme RNA polymerase. This enzyme synthesizes RNA in a test tube from a mixture of the A, U, C, and G bases, but it does so only in the presence of a primer DNA. The sequence of the bases in the primer is copied in the RNA. The steps involved in this process are as follows: (1) the DNA double helix unwinds by breaking the hydrogen bonds between the corresponding bases in the paired strands; (2) the RNA polymerase forms the bonds between the RNA bases that are complementary to the bases in the DNA; and (3) the messenger RNA thus formed passes into the cytoplasm and becomes attached to a ribosome. Ribosomes consist of proteins and another type of RNA, ribosomal RNA (rRNA).

Protein synthesis: translation. The process of protein synthesis is represented diagrammatically in Figure 10. The information contained in the sequence of the bases (letters) in the messenger RNA is then translated into a sequence of amino acids in a protein. This requires the presence of still another molecule that is capable of recognizing the code for a specific amino acid and selectively making the amino acid available at the right point in the protein synthesis, a soluble RNA fraction within cells that can bind amino acids. Soluble, or transfer, RNA (sRNA, or tRNA) is a single-stranded molecule that forms about 20 percent of the total cellular RNA. If amino acids and a source of energy (usually ATP) are added to a mixture of transfer RNA's, reversible binding of the amino acids to the RNA molecules occurs. Furthermore, each amino acid is bonded to a specific transfer RNA molecule by a specific activating enzyme. There are at least 20 different kinds of transfer RNA's and activating enzymes that correspond to the 20 amino acids commonly found in proteins. The amino acid-transfer RNA complex becomes attached to the ribosome with its messenger RNA molecule; the addition of the amino acid to the growing polypeptide chain then occurs. A sequence of three nitrogenous bases (anticodon) on the transfer RNA molecule pairs with a complementary sequence (codon) on the messenger RNA molecule, which is held in the correct position by the ribosome. Once the recognition has occurred, a peptide bond is formed between the amino acid bound to the transfer RNA and the growing polypeptide chain.

From T. Jukes *Molecules and Evolution* (1966), Columbia University Press

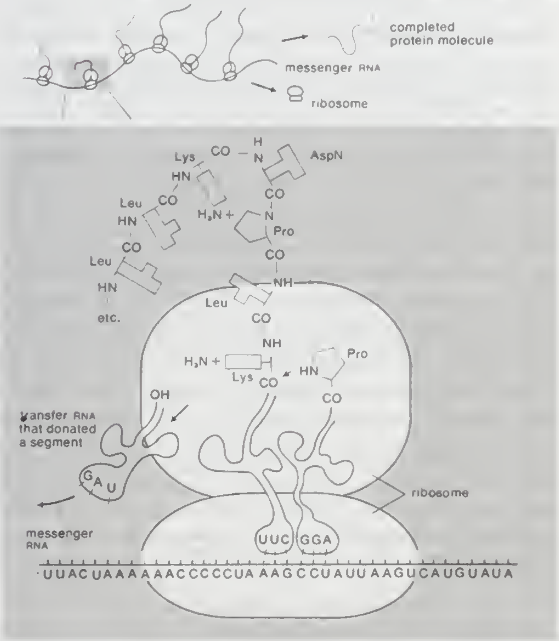


Figure 10: *Synthesis of protein.*
(Top) The messenger RNA shown with a ribosome (in stages) "reading" its message, the translation occurring from left to right, and the polypeptide chain growing as the ribosome proceeds along the messenger RNA. (Bottom) A highly magnified ribosome in the process of "reading" a messenger RNA molecule. The ribosome has just "read" an AAG sequence off the messenger RNA, and a specific transfer RNA molecule (UUC) moves into place carrying the amino acid lysine (Lys), which is then added to the polypeptide chain. Next in position to be read is CCU, which draws a transfer RNA molecule carrying the amino acid proline (Pro) to be transferred to the growing polypeptide chain. The ribosome moves from left to right along the messenger RNA molecule. Each ribosome has two binding sites for transfer RNA: one that holds the growing polypeptide chain and a second for positioning a new transfer RNA molecule for processing.

The accuracy of the model described in Figure 10 has been confirmed by the achievement of protein synthesis in the test tube. This synthesis requires a DNA template (primer DNA), precursor nucleotide molecules, ribosomes, transfer RNA's, amino acids, and a set of enzymes and certain other factors.

The central dogma. The processes of transcription and translation, as described above, can be represented thus: DNA → RNA → protein. Soon after its elucidation, this understanding of the genetic control of protein synthesis became known as "the central dogma" of molecular genetics. Included as part of the dogma was the belief that reverse transfer of information does not occur; in other words, there is no storage of information in the protein molecules and no transcription of protein back into nucleic acids or of RNA back into DNA.

The central dogma has since been modified to accommodate the discovery that reverse transcription of RNA to DNA does occur, as first demonstrated in some viruses. These viruses, called retroviruses, have a genome composed of RNA. When retroviruses enter a host cell, they produce an enzyme called reverse transcriptase. This enzyme permits the transcription of the viral RNA into DNA, which then may become incorporated into the genetic material of the host cell.

A second modification was necessitated by the discovery that not all DNA codes for protein synthesis. As discussed below, some of the noncoding DNA is involved in regulating the biochemical processes of the cell. The amount of noncoding DNA is small in procaryotes, but in eucaryotes it may be most of the cell's DNA.

The words of the genetic code

Reading the code. It is necessary to understand how the four letters—A, T, C, and G—specify, or code, for 20 different amino acids. If a single letter coded for an amino acid, only four amino acids could be specified. If two bases were needed to specify an amino acid, then 16 different combinations could be constructed, again an insufficient number (20 amino acids must be accounted for). Combinations of three letters allow 64 different words to be constructed, more than the necessary minimal number. A three-letter, triplet, code could be constructed in at least three different ways: (1) with words overlapping; (2) with words not overlapping and punctuated; and (3) with words not overlapping and not punctuated. An overlapping code is composed of words that overlap each other—*i.e.*, the letters of any given word may belong to one, two, or three words. The DNA might contain, for example, the sequence

$\underline{A\,G\,\overline{C}}\,\overline{G}\,\underline{\overline{T}\,T\,\overline{A}}\,\overline{C\,G}$; the first word is AGC, the second CGT, and so on. This type of code is improbable, because of the restrictions it would place upon the possible sequence of amino acids in protein. As the example above shows, if the first word is AGC, the second word must begin with C, etc. Examination of amino-acid sequences in a protein such as hemoglobin indicates that any amino acid can follow any other—a possibility not allowed for by an overlapping code.

If the code is nonoverlapping, a problem of distinguishing words from each other arises. DNA contains no spaces separating the words as in written sentences; therefore, there must be other indications of specific starting points for messenger RNA synthesis. The base sequence AGC AGC AGC . . . could be punctuated by the presence of a fourth base, T, between each AGC triplet. This would reduce the number of possible triplets to 27. That a punctuated code of this type is not realized is seen from the evidence of the degeneracy in the code for some amino acids. The degeneracy means that some amino acids are coded for by more than one triplet, and a punctuated code does not allow enough words. A second objection to this type of code comes from a consideration of the effects of mutation on the coding sequence. If one of the punctuation marks mutates to another base, or a coding base mutates to a punctuation mark, the resulting sequence will be complete nonsense functionally.

The third possibility is a nonoverlapping, nonpunctuated code, in which the reading starts from a specific point. In all organisms studied in this respect this is the method of coding used. A knowledge of the base sequence in the messenger RNA and the resulting amino-acid sequence in protein reveals the code for each amino acid. The triplet UUU, for example, is the code for the amino acid phenylalanine, corresponding to the sequence AAA in the DNA. Poly-A (AAA) and poly-C (CCC) are messenger RNA's codes for lysine and proline, respectively.

Other triplets were tested for their coding abilities by synthesizing messenger RNA molecules with varying proportions of the two bases. If, for example, a mixture of the two bases U and C in a 5 : 1 proportion are synthesized into RNA, the possible triplets and their probable frequency in the synthetic messenger RNA can be easily determined. The triplet UUU will be most common and will appear with the frequency 5/6 × 5/6 × 5/6; the triplets UUC, UCU, and CUU will appear in the frequencies of 5/6 × 5/6 × 1/6; the triplets UCC, CUC, and CCU will be the next most frequent and will appear with a frequency of 5/6 × 1/6 × 1/6; while the triplet CCC should appear only 1/216 of the time. A messenger RNA of this composition should result in the incorporation into protein of eight different amino acids. In fact, only four amino acids were present in the protein produced; this means that several of these triplets encode for the same amino acid and therefore that the code is degenerate.

The RNA code triplets (or codons) and the amino acids for which they stand are shown in Table 3. Triplets have been discovered that encode for starting and for stopping the synthesis of protein chains in *E. coli.* Many proteins of *E. coli* begin with the amino acid methionine. Two different transfer RNA's for methionine are known to exist, only one of which functions to initiate protein synthesis. After synthesis of the protein, an enzyme may remove a portion of the beginning of the chain to eliminate the obligatory methionine molecule. The second transfer RNA for methionine allows this amino acid to be incorporated into the middle of a polypeptide.

Termination of the synthesis of a polypeptide chain is signalled by three different RNA codons that do not specify an amino acid: UAA, UAG, and UGA. These triplets were discovered as nonsense mutations that produced premature cessation of protein synthesis in many different genes. Specific proteins called release factors can read these codons and release the polypeptide chain from the ribosome.

MUTATIONS IN THE CODE

The DNA content of the cell must accurately replicate itself prior to mitosis or meiosis. Given the complexity of the DNA molecule and the vast number of cell divisions that take place within the lifetime of a multicellular organism, it is obvious that copying errors are likely to occur. If unrepaired, such errors change the linear order of the DNA bases and produce mutations in the genetic code. Many mutations arise from unknown causes. In addition to these so-called spontaneous mutations, researchers have demonstrated that a variety of environmental agents—including ionizing radiation, toxic chemicals, and certain viruses—can induce mutations. The effects of these mutagenic agents on human health are discussed below in the section *Human genetics.*

Table 3: The Genetic Code: Nucleotide Triplets (Codons) Specifying Different Amino Acids in Protein Chains*

DNA triplet	RNA triplet	amino acid	DNA triplet	RNA triplet	amino acid
AAA	UUU	phenylalanine	ACA	UGU	cysteine
AAG	UUC		ACG	UGC	
AAT	UUA	leucine	ACC	UGG	tryptophan
AAC	UUG		ATA	UAU	tyrosine
GAA	CUU		ATG	UAC	
GAG	CUC		ATT	UAA	(termination: end of specification)
GAT	CUA		ATC	UAG	
GAC	CUG		ACT	UGA	
AGA	UCU	serine	GCA	CGU	arginine
AGG	UCC		GCG	CGC	
AGT	UCA		GCT	CGA	
AGC	UCG		GCC	CCG	
TCA	AGU		TCT	AGA	
TCG	AGC		TCC	AGG	
GGA	CCU	proline	GTA	CAU	histidine
GGG	CCC		GTG	CAC	
GGT	CCA		GTT	CAA	glutamine (GluN)
GGC	CCG		GTC	CAG	
TAA	AUU	isoleucine (Ileu)	TTA	AAU	asparagine (AspN)
TAG	AUC		TTG	AAC	
TAT	AUA		TTT	AAA	lysine
TAC	AUG	methionine	TTC	AAG	
TGA	ACU	threonine	CCA	GGU	glycine
TGG	ACC		CCG	GGC	
TGT	ACA		CCT	GGA	
TGC	ACG		CCC	GGG	
CAA	GUU	valine	CTA	GAU	aspartic acid
CAG	GUC		CTG	GAC	
CAT	GUA		CTT	GAA	glutamic acid
CAC	GUG		CTC	GAG	
CGA	GCU	alanine			
CGG	GCC				
CGT	GCA				
CGC	GCG				

* The columns may be read: the DNA triplet is transcribed into an RNA triplet, which then directs the production of an amino acid.

Kinds of mutations. The addition or deletion of one or more bases results in a frame-shift mutation, so named because the reading frame of the gene, and thus its message, is altered from that point forward. Suppose that a DNA message read from left to right reveals the triplets GAC, TCA, and TTA (which are transcribed in the RNA code as CUG, AGU, and AAU). Deletion of the first T alters the reading frame so that triplets GAC, CAT, TA . . . will be read. The first triplet is unchanged, but all the remaining triplets may specify wrong amino acids. The chemical addition of a base to the sequence likewise shifts the reading frame; such a mutant will also specify wrong amino acids beyond the point of the base addition. If an original mutant resulted from the deletion of a base from the DNA, the addition of a base at a point beyond the first mutation would restore the reading frame of the DNA sequence and would result in nearly normal function. For example, assume that the original DNA sequence reads ACT GGC TAG CTG TCA TCG . . . . Deletion of the C in the second triplet results in the following triplets being read: ACT, GGT, AGC, TGT, CAT, CG . . . . The subsequent addition of a base (A) between the third and fourth mutant triplets results in the following sequence: ACT GGT AGC ATG TCA TCG . . . . Note that the first, fifth, and sixth triplets are identical to those in the original sequence. Only the second, third, and fourth triplets are altered, and the reading of the code from the fifth triplet on will be identical to that in the original message. Frequently, suppressor mutations occur in proximity to other mutations and restore the reading frame of the DNA sequence, thereby allowing a sequence of amino acids differing only slightly from the original one to be formed in the protein.

Mutations in which one base is exchanged for another are called base substitutions, or point mutations. A base substitution may result in the incorporation of one wrong amino acid into the polypeptide chain encoded by the gene. What effect this has on the functioning of the protein of which the chain is part depends on the type and position of the wrong amino acid. In many cases, the effects are minor, but there are exceptions. The human disease sickle-cell anemia, for example, is the product of a single base substitution inherited from both parents. Sometimes a base substitution results in a codon for an amino acid being changed to one of the termination triplets. This type of point mutation will cause premature termination of protein synthesis and, probably, complete loss of function in the finished protein.

Thus far distinctions have been made between mutations in terms of their effects on the nucleotide sequence of DNA. It is also useful to differentiate between mutations that affect germ cells (*i.e.*, eggs and sperm) and those that affect somatic cells. When a mutation occurs in a germ cell, it can be passed on to offspring, where it will be carried in every cell of the new individual. Mutations in somatic cells, on the other hand, are not passed on to offspring, and they affect only a certain population of cells (the original mutant cell and its mitotic descendants) within the affected individual. (Ro.R./T.D./Ar.R.)

Mutation rates. Detectable results of germinal mutation among people are only very rarely encountered. Thus, the actual rate of mutation in human chromosomes defies full measurement. A major reason for this is that most mutations seem to be recessive and thus tend to be masked for generations. Efforts to measure mutation rate therefore are most conveniently directed toward selected dominant or codominant mutations for which phenotypic recognition is easier. Indirect (inferential) methods of measurement are still required.

Studies of achondroplasia

One dominant gene that is useful for studying human mutation rates produces the form of dwarfism called achondroplasia. When an affected child appears in a family in which both parents are normal, the properly diagnosed condition can be ascribed to the occurrence of one new mutation. The frequency of such an event is customarily calculated on the basis of the number of gametes (egg and sperm cells) produced by the parents in one generation; for human achondroplasia, the mutation rate has been inferred to be 4.2 per 100,000 gametes. Analyses of a number of different gene loci in humans and in such experimental organisms as corn and the *Drosophila* fly show that the average mutation rate among living beings is on the order of one in 100,000 gametes (10^{-5}). Nevertheless, each gene studied shows its unique mutational probability; the neurological disorder called Huntington's chorea shows only about 0.5 mutation per 100,000 gametes, whereas the figure for neurofibromatosis (a disorder with soft tumours distributed over the whole body) has been stated to be somewhat higher. The general average in humans is considered to be about the same as for achondroplasia, roughly four per 100,000 gametes. (H.L.C.)

It may well be, however, that mutation rates are considerably higher than this figure for the following reasons: (1) there undoubtedly are "silent" mutations that do not change the biological function of the gene product (protein) in a way that will change the phenotype; (2) some mutations may be so harmful as to be lethal early in embryonic development; and (3) different mutations can produce the same abnormal phenotype, a situation known as genetic heterogeneity. It follows from the above that more accurate mutation rates in humans will result only from DNA analysis that reveals changes in the specific nucleotides of the DNA chain.

DNA repair. A variety of agents in the cell's environment, both chemical and physical, can damage DNA. Organisms have developed a variety of mechanisms for repairing copying errors produced by damaged DNA, usually by enzymatically excising them. The enzyme DNA polymerase then catalyzes the replacement of the excised segment with the correct nucleotides, using the undamaged DNA strand as a template. Eucaryotic cells have a greater variety of DNA repair mechanisms than do bacteria. Malfunctioning of the repair mechanisms can lead to genetic disease, abnormal function, or cancer. Xeroderma pigmentosum, a lethal human disease that is recessively inherited, involves several defective repair mechanisms.

REGULATION OF GENES

The evidence accumulated in genetics makes it virtually certain that not all the genes present in a cell are active in directing the specific processes of protein synthesis. Gene action can be switched on or off in response to the cell's stage of development and external environment. In multicellular organisms, moreover, each cell has a complete copy of the organism's genetic instructions, though different kinds of cells come to express different parts of the genome. That is to say that a skin cell, a nerve cell, and a bone marrow cell from the same person all contain the same genes; the differences in structure and function among these cells result from the selective expression and repression of certain genes.

In bacteria. In 1961 the French molecular biologists François Jacob and Jacques Monod proposed a model for genetic regulation in *E. coli*. When grown on a minimum culture of carbohydrates and sulfur, these bacteria can synthesize all of their necessary amino acids. To accomplish this amino-acid synthesis, the bacteria must produce various enzymes, the activities of which can be detected in a growing culture. If certain amino acids are added to the culture medium, the bacteria stop producing the enzymes required for the synthesis of these amino acids. This phenomenon is known as repression, and the enzymes affected are repressible enzymes. The pathway for the synthesis of the amino acid arginine in *E. coli* is a good example of how these repressible enzymes work. This synthesis involves three steps and three separate enzymes:

$$\text{glutamic acid} \xrightarrow{\text{enzyme A}} \text{ornithine} \xrightarrow{\text{enzyme B}} \text{citrulline} \xrightarrow{\text{enzyme C}} \text{arginine.}$$

When arginine is present in the medium, none of the three enzymes involved in this process is detected, but if arginine is removed, all three enzymes rapidly appear. The end product of this pathway, arginine, controls the production of the intermediate enzymes, since the addition of either ornithine or citrulline to the medium has no effect. A related process involves the production of enzymes whose substrates are not always present in cells. The presence of lactose in the medium, for example, induces the synthesis of three enzymes that procced to degrade lactose; this phenomenon is termed induction.

The classes of genes involved in regulating the expression of a bacterial gene—repressing its action or inducing it—are shown in Figure 11. The part of the chromosome containing the genes concerned is divided into two regions, one of which includes the structural genes (*i.e.*, those genes that code for protein synthesis) and the operator gene. This region is termed an operon. The other part contains only the regulator gene. The regulator gene need not be located close to the operon. The regulator gene produces some substance, a repressor, which affects a second gene, an operator. There are several lines of evidence that suggest that the repressor substance is a protein molecule. It would be necessary for such a molecule to have at least two areas or sites that interact with the operator gene or a metabolite or both in order to influence structural genes (repress or induce their action, depending on given conditions).

From F Jacob and J Monod, *Journal of Molecular Biology* (1961), Academic Press Inc

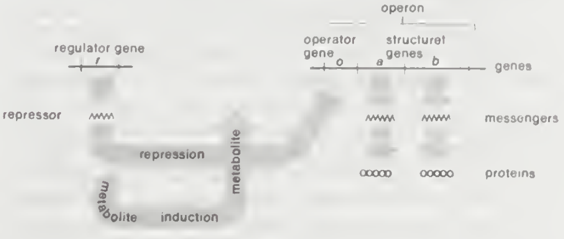


Figure 11: Model of the operon and its relation to the regulator gene.

In higher organisms. Although the operon model has proved a useful description of gene regulation in bacteria, evidence indicates that different regulatory mechanisms are employed in eucaryotes. The series of events associated with gene expression in higher organisms is much more complex than in procaryotes, requiring multiple levels of regulation. This is particularly true in multicellular organisms, in which cell differentiation takes place.

The regulation of gene expression in eucaryotes is not fully understood, but research in the fields of somatic cell genetics and recombinant DNA have yielded at least partial explanations as to how this process occurs. (The techniques involved in such research are discussed below.) What this research has indicated may be the mechanisms of gene regulation in higher organisms.

Eucaryotic DNA comprises three different classes: (1) unique, or single copy, DNA, which contains the structural genes (protein coding sequences); (2) moderately repetitive DNA, some forms (families) of which are dispersed throughout the chromosomes in small clusters and which may contain some functional genes, such as those that code for certain forms of RNA; and (3) the highly repetitive DNA, which contains nucleotide sequences repeated up to 1,000,000 times. This last class of DNA is usually clustered near the centromeres and is often known as satellite DNA. The function of the families of satellite DNA is unknown. There is no indication that their remarkably constant sequences are ever transcribed.

Since a very small percent of the DNA in higher organisms codes for proteins, the assumption is that the majority of DNA is involved in the control of gene action. If all the human structural genes functioned simultaneously, for example, metabolic chaos would ensue. Mechanisms exist for switching gene activity on and off at the appropriate time in development and in appropriate tissues, thus permitting the differentiation of the organs at various stages of development. Much of this regulation in higher organisms occurs during the processing of the RNA copies transcribed from DNA. Following the transcription of the DNA into this nuclear, or heterogenous, RNA, a process of editing and splicing takes place. The nuclear RNA contains long and often multiple noncoding intervening sequences of nucleotides. These sequences, called introns, are cut out, and the remaining coding regions, called exons, are spliced together to form the functional messenger RNA that can leave the nucleus to direct protein synthesis in the cytoplasm (see Figure 12). Because the nuclear RNA, including its introns, is a mirror image of its DNA template, it follows that many genes of eucaryotes are discontinuous and may be characterized as mosaics of coding and noncoding regions. The role of the introns is not firmly established, but some evidence points to their possible involvement in regulating gene action. For example, researchers have shown that several forms of thalassemia (a common disease of hemoglobin) stem from intron mutations that interfere with the splicing action.

With the discovery of introns, molecular geneticists realized that it is impossible to determine the nucleotide sequence of a eucaryotic structural gene simply by analyzing the amino-acid sequence of its polypeptide product. This is so because the amino-acid sequence will not reflect the introns present in the native DNA. What is possible is to reconstruct the complementary DNA (CDNA), the DNA sequences consisting only of exons. As will be seen later, this has become an important part of recombinant DNA technology.

The analysis of nucleotide sequences has revealed a number of other mechanisms that help in the regulation of gene expression. Molecular geneticists have discovered small DNA sequences, about 100 base pairs in length, that interact with other constituents of the cell to produce specific regulatory effects on adjacent genes. These regulatory sites—called promoters, enhancers, or modulators—have been sequenced, and their effects studied on the quantitative expression of the products of a "standard" gene cloned by recombinant DNA techniques. Of particular interest is the enhancer site which may be involved in turning on specific genes in specific tissues at specific stages in development. In addition to illuminating the mysterious problems associated with normal development, chromosomal translocations may move these sites to abnormal positions where they can disturb cell multiplication and produce malignancy.

CYTOPLASMIC, OR EXTRANUCLEAR, DNA

Some characteristics of organisms do not show Mendelian segregation and, among higher organisms, are inherited through the maternal line only. For example, the cells of green plants contain cytoplasmic bodies called chloroplasts, which carry the green pigment chlorophyll. In corn, variants are known whose leaves are striped green and yellow because of the absence of chlorophyll in the chloroplasts of some cells. Since no chloroplasts are present in pollen grains, the conclusion is that chloroplasts are self-replicating bodies in egg cells and in part control their own characteristics.

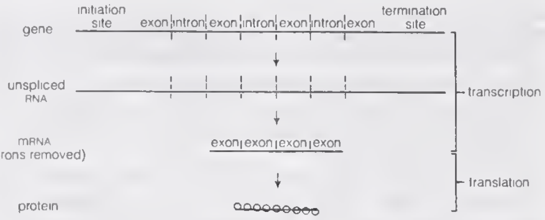


Figure 12: The steps in the processing of a gene in eucaryotes (see text).

In the 1960s the existence of DNA within chloroplasts (in plants) and mitochondria (in complex animal cells) was demonstrated. The subsequent discoveries of ribosomes, transfer RNA's, and enzymes within mitochondria and chloroplasts have demonstrated that these cytoplasmic bodies synthesize part, but not all, of their own proteins.

The usual procedure for testing for extranuclear inheritance is to look at the progeny of reciprocal crosses of two different strains. Chromosomal (nuclear) inheritance predicts that offspring of the cross male A × female B should not differ phenotypically from those of the cross male B × female A. Cytoplasmic inheritance, on the other hand, is entirely maternal, as only the female gamete contributes chloroplasts or mitochondria to the zygote. Hence the offspring of male B × female A differ from those of male A × female B for traits controlled by extranuclear DNA. There is little evidence of cytoplasmic inheritance in humans. A large pedigree, however, has been reported of a family with a disease characterized by abnormal muscle fibres and abnormal mitochondria. In every case the disease was inherited from an affected mother, suggesting the presence of mutant mitochondrial DNA and cytoplasmic inheritance.

THE BIOCHEMICAL PRODUCTS OF GENE EXPRESSION

As stated above, the central function of genes (and hence of DNA) is to direct the production of proteins. The conceptual union of biochemistry and genetics, now called biochemical genetics, was first envisaged by Garrod, an English physician who has been called "the father of chemical genetics." In lectures delivered in 1908, Garrod described four hereditary diseases (alkaptonuria, cystinuria, albinism, and pentosuria) that involve an enzyme deficiency. He dubbed these diseases "inborn errors of metabolism," a name that persists to this day. Garrod stressed the unusual frequency of consanguineous matings in the parents of these patients. At a time when Mendel's laws were just being rediscovered, Garrod gave the first detailed description of recessive inheritance in humans, a point quickly appreciated by a pioneer in genetics, William Bateson.

Although the relationship between gene and enzyme was implicit in Garrod's work, it was more than 30 years before George W. Beadle and Edward L. Tatum of the United States made it explicit by their classical studies on genetic control of biochemical reactions in the bread mold *Neurospora crassa.* This work, for which Beadle and Tatum received the Nobel Prize for Physiology or Medicine for 1958, resulted in the formulation of the "one gene–one enzyme" hypothesis.

Most strains of *N. crassa* produce enzymes that enable the mold to grow on a minimal medium of sugar, inorganic salts, and the vitamin biotin. From these sources, *N. crassa* can synthesize all the amino acids, vitamins, and other substances necessary for its survival and reproduction. If, however, a genetic mutation results in the production of a chemically defective enzyme, the affected strain will not grow on the minimal medium unless the product of the now-defective enzymatic reaction is added. Let a metabolic pathway consist of a chain of reactions $A \xrightarrow{(1)} B \xrightarrow{(2)} C \xrightarrow{(3)} D$, in which each reaction is mediated by a specific enzyme (1, 2, and 3, respectively). If the gene for enzyme 2 is mutated, C cannot be synthesized and the reaction stops. If, however, C is added to the medium, the reaction proceeds and D is formed. Meanwhile, however, B will accumulate in large amounts, which may be toxic.

The existence of metabolic pathways composed of successive steps, each of which is controlled by a single gene, appears to be a general phenomenon in the living world. One of the experimental techniques of somatic cell genetics (see below) has been to produce single-gene mutations that cause a nutritional defect in the experimental cell line. Through these abnormalities geneticists can elucidate the normal metabolic pathways of the cell and can map the gene to a specific chromosomal site. A similar phenomenon, of course, operates in the inborn errors of metabolism. Well over 100 different inborn errors of human metabolism are understood as defects in specific enzymes, secondary to specific gene mutations.

It was not, however, until the work of the U.S. biochemist Linus Pauling and his associates (1949) that the molecular basis for the relationship between a mutant gene and its protein product began to be understood. This was accomplished by the use of electrophoresis, a technique that separates different proteins by their movement through a liquid under the influence of an electric field. Pauling's electrophoretic separation of normal hemoglobin (hemoglobin A) from the hemoglobin of sickle-cell anemia (hemoglobin S) made it apparent for the first time that genetic disease could be understood in molecular terms. The hemoglobin molecule consists of two pairs of polypeptide chains, an alpha chain and a beta chain. Of fundamental importance was the determination by Vernon M. Ingram that the electrophoretic difference between hemoglobins A and S is due to the change of a single amino acid—glutamic acid being replaced by valine in position six of the beta chain. This results from a single base change in the DNA triplet responsible for the amino acid in position six: GAA is changed to GUA. The beta chain has 146 amino acids, but this one change disturbs the structure of the hemoglobin sufficiently to produce a most serious disease, sickle-cell anemia, in the homozygote.

Information resulting from this and other hemoglobin diseases demands that the "one gene–one enzyme" hypothesis be modified to the "one gene–one polypeptide" hypothesis or even more accurately to the "one structural gene–one polypeptide" principle. This has the advantage also of broadening the concept of inborn errors of metabolism to include any inherited deficiency of a protein, whether the protein is involved in enzymatic activity, transport, or structure. (Ro.R./T.D./Ar.R.)

SOMATIC CELL GENETICS

The development of molecular biology represented a fusion of biochemistry and genetics. As has been discussed, most of the pioneer research in this field utilized microorganisms. The great strides made in genetic-biochemical analysis resulted basically from the ability to place an experimental organism on a culture dish containing agar (a jellylike substance) and a nutrient medium supporting cell multiplication. The cells multiplied to produce discrete colonies. All the cells in a particular colony formed a clone; that is, they all had the same genetic constitution as the founding cell. By selecting a founding cell with a phenotypically observable mutation, researchers could easily establish clones with the mutant genotype. The biochemical products of these mutant cells could then be studied and compared to those of normal colonies.

By contrast, the study of genetics in higher organisms was, for many years, limited to the analysis of the genetic effects of experimental matings. In human genetics, even this technique was not available, as the experimental mating of human beings is not only morally unjustifiable but also unworkable due to the long generation time (about 30 years) of the species. These obstacles precluded the observation of the segregation of human genes over multiple generations except by the classical methods of retrospective analysis of large family pedigrees.

During the late 1950s, molecular biologists learned to grow single mammalian somatic cells in culture, a feat hitherto thought impossible. This ability to treat "the mammalian cell as a microorganism" (a phrase coined by the U.S. molecular biologist Theodore T. Puck) has made it possible to study the genetics of higher organisms with techniques similar to those used in *E. coli.* Somatic cell genetics has permitted researchers to analyze mammalian cells in terms of their growth requirements and their responses to a variety of environmental agents and especially to study, at the molecular and cellular level, the processes of heredity in these cells. Because of such research, the characteristics of mammalian cells in culture can be analyzed in simple, quantifiable terms. In the study of human genetics, somatic cell techniques have revolutionized the diagnosis of genetic disease (including prenatal diagnosis), have made possible the analysis of the human chromosomes, and have elucidated some of the processes of normal and abnormal differentiation, including those involved in producing cancer.

Mammalian cells can be cultured from many different tissues (*e.g.,* white cells from the blood, fibroblasts from the skin, marrow cells from the bone) for a variety of purposes. These include (1) studies of the chromosomes (cytogenetics), (2) biochemical assays to look for enzymatic defects associated with human diseases, (3) examination of DNA and the processes of replication, (4) interspecific and intraspecific somatic cell hybridization to gain information about the position of genes on the human chromosomes, and (5) the study of the structure and function of the cells and their organelles (constituent parts). These cells, with the exception of lymphocytes (white blood cells), can be kept in culture for long periods or stored in a frozen state to be thawed out and recultured later. The human lymphocyte, readily available from a peripheral blood sample, can be kept in culture for only several days. It can, however, be transformed into a permanent culture that can be frozen and reactivated at a later time by infecting it with certain viruses, particularly the Epstein–Barr virus.

The ability to fuse two cells, either of the same or of different mammalian species, has proved to be one of the most powerful tools of somatic cell genetics. Researchers accomplish such fusion by treating the cultured cells with an irradiated, killed virus (the sendai virus) or with particular chemical agents; in response the cytoplasm and nuclei of the cells merge to form a hybrid cell with a single, large nucleus that contains the genome of both "parent" cells.

The hybridization of two cells from the same species—each cell having been exposed to radiation or chemical agents to produce a desired phenotypic mutation—is useful in complementation analysis. Consider, for example, the fusion of two mutant cells (A and B) whose DNA has been damaged by a point mutation so that neither cell can grow in a medium lacking glycine. If the hybrid cell can grow in the absence of glycine, the two mutant genomes have complemented each other. In short, the defect in the DNA present in cell A involves a different gene at a different location than that in cell B, and the combined genome is able to compensate for each mutation. If, however, A and B have mutations at the same DNA locus, they cannot produce a glycine-independent hybrid. When complementation occurs, researchers can analyze the biochemical products of the "parent" cells to elucidate the nature of the mutation that exists in each cell.

The fusion of cells of two different species (*e.g.,* of a human cell and a Chinese hamster cell) produces a hybrid cell that, when cultured, undergoes extensive chromosome loss, primarily of the species whose cultured cells have the longer generation time. In the case of the human–Chinese hamster hybrid, human chromosomes are lost. If a mutant hamster cell with a nutritional deficiency (*e.g.,* one that cannot grow in the absence of glycine) is hybridized with a human cell without the deficiency, the hybrid cell will not show the deficiency (in this instance the hybrid will be able to grow in a medium without glycine). The hybrid cells will then produce clones showing extensive loss of the human chromosomes. The one human chromosome the hybrid cannot lose is the one that carries the human gene that compensates for the hamster cell mutation, for without that gene the cell would die. By employing this technique, researchers can identify the human chromosome on which a particular gene is located. In fact, by using a series of X rays or other manipulations to break the human chromosome, geneticists can locate the segment of the chromosome possessing the gene. Hundreds of human genes have been mapped to a specific chromosome, in many cases to a specific chromosomal region, by this method. Using increasingly small chromosome deletions, researchers have succeeded in locating the genes in their proper order on the chromosomes. By studying the recombination frequency between these genes (a task made simpler by the methods of recombinant DNA discussed below), the distances between them on the chromosome have been mapped. The ability to locate the human genes on the 23 pairs of chromosomes is of tremendous practical importance for understanding human genetics. When large numbers of genes are mapped, investigators can identify which are absent or supernumerary in the many diseases involving the human chromosomes. Mapping also helps geneticists elucidate the nature of gene control in mammals, which, as has been discussed, is significantly more complex than the relatively simple operon system in bacteria. Insights into gene regulation in mammals can increasingly be applied to understanding of cellular biochemical control in normal development and in human disease.

RECOMBINANT DNA TECHNOLOGY

Since the focus of all genetics is the gene, the fundamental goal of geneticists is to isolate, characterize, and manipulate genes. Although it is relatively easy to isolate a sample of DNA from a collection of cells, finding a specific gene within this DNA sample can be compared to finding a needle in a haystack. Consider the fact that each human cell contains approximately two metres (approximately six feet) of DNA. Therefore, a small tissue sample will contain many kilometres of DNA. However, recombinant DNA technology has made it possible to isolate one gene or any other segment of DNA, enabling researchers to determine its nucleotide sequence, study its transcripts, mutate it in highly specific ways, and reinsert the modified sequence into a living organism.

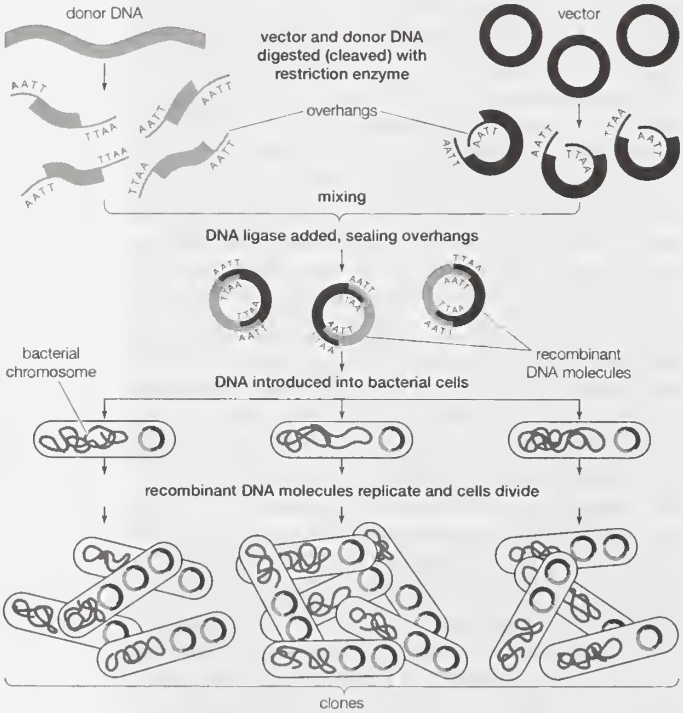

donor DNA
vector
vector and donor DNA digested (cleaved) with restriction enzyme
AATT
TTAA
overhangs
mixing
DNA ligase added, sealing overhangs
bacterial chromosome
recombinant DNA molecules
DNA introduced into bacterial cells
recombinant DNA molecules replicate and cells divide
clones

Figure 13: Steps involved in the production of a recombinant DNA molecule.
Encyclopædia Britannica, Inc.

DNA CLONING

In biology, a clone is a group of individual cells or organisms all descended from one progenitor. This means that the members of a clone are genetically identical because cell replication produces identical daughter cells each time. The use of the word *clone* has been extended to recombinant DNA technology. If it were possible to produce many copies of a single fragment of DNA, such as a gene, these identical copies would constitute a DNA clone. Indeed, this is now possible. In practice, the procedure is carried out by inserting a DNA fragment into a small DNA molecule and then allowing this molecule to replicate inside a simple living cell such as a bacterium. The small replicating molecule is called a DNA vector (carrier). The most commonly used vectors are plasmids, circular DNA molecules that originated from bacteria, viruses, and yeast cells. Plasmids are not a part of the main cellular genome, but they can carry genes that provide the host cell with useful properties such as drug resistance, mating ability, and toxin production. They are small enough to be conveniently manipulated experimentally, and furthermore they will carry extra DNA that is spliced into them. The process of gene cloning is illustrated in Figure 13.

Restriction enzymes

Creating the clone. The steps in cloning are as follows. DNA is extracted from the organism under study and is cut into small fragments of a size suitable for cloning. Most often this is achieved by cleaving the DNA with a restriction enzyme. Restriction enzymes are extracted from several different species and strains of bacteria, in which they act as defense mechanisms against viruses. They can be thought of as "molecular scissors," cutting the DNA at specific target sequences. The most useful restriction enzymes make staggered cuts; that is, they leave a single-stranded overhang at the site of cleavage. These overhangs are very useful in cloning because the unpaired nucleotides will pair with other overhangs made using the same restriction enzyme. So, if the donor DNA and the vector DNA are both cut with the same enzyme, there is a strong possibility that the donor fragments and the cut vector will splice together because of the complementary overhangs. The resulting molecule is called recombinant DNA. It is recombinant in the sense that it is composed of DNA from two different sources. Thus, it is a type of DNA that would be impossible naturally and is an artifact created by DNA technology.

The next step in the cloning process is to cut the vector with the same restriction enzyme used to cut the donor DNA. Vectors have target sites for many different restriction enzymes, but the most convenient ones are those that occur only once in the vector molecule. This is because the restriction enzyme then merely opens up the vector ring, creating a space for the insertion of the donor DNA segment. Cut vector DNA and donor DNA are mixed in a test tube, and the complementary ends of both types of DNA unite randomly. Of course, several types of unions are possible: donor fragment to donor fragment, vector fragment to vector fragment, and, most important, vector fragment to donor fragment. There are special ways of selecting for vector molecules with donor inserts, as described below. Recombinant DNA associations form spontaneously in the above manner, but these associations are not stable because, although the ends are paired, the sugar-phosphate backbone of the DNA has not been sealed. This is accomplished by the application of an enzyme called DNA ligase, which seals the two segments, forming a continuous and stable double helix.

The mixture should now contain a population of vectors each containing a different donor insert. This solution is mixed with live bacterial cells that have been specially treated to make their cells more permeable to DNA. Recombinant molecules enter living cells in a process called transformation. Usually, only a single recombinant molecule will enter any individual bacterial cell. Once inside, the recombinant DNA molecule replicates like any other plasmid DNA molecule, and many copies are subsequently produced. Furthermore, when the bacterial cell divides, all of the daughter cells receive the recombinant plasmid, which again replicates in each daughter cell.

The original mixture of transformed bacterial cells is spread out on a surface of a growth medium in a flat dish

(Petri dish), so that the cells are separated from one another. These individual cells are invisible to the naked eye, but as each cell undergoes successive rounds of cell division, visible colonies form. Each colony is a cell clone, but it is also a DNA clone because the recombinant vector has now been amplified by replication during every round of cell division. Thus, the Petri dish, which may contain many hundreds of distinct colonies, represents a large number of clones of different DNA fragments. This collection of clones is called a DNA library. By considering the size of the donor genome and the average size of the inserts in the recombinant DNA molecule, a researcher can calculate the number of clones needed to encompass the entire donor genome, or, in other words, the number of clones needed to constitute a genomic library.

Types of genetic libraries

Another type of library is a complementary DNA (cDNA) library. Creation of a cDNA library begins with mRNA instead of DNA. These mRNA molecules are treated with the enzyme reverse transcriptase, which is used to make a DNA copy of an mRNA. The resulting DNA molecules are called cDNA. A cDNA library represents a sampling of the transcribed genes, whereas a genomic library includes untranscribed regions.

Both genomic and cDNA libraries are made without regard to obtaining functional cloned donor fragments. Genomic clones do not necessarily contain full-length copies of genes. Furthermore, genomic DNA from eukaryotes contains introns, which cannot be processed by bacterial cells, so even full-sized genes would not be translated. In addition, eukaryotic regulatory signals are different from those used by prokaryotes. However, it is possible to produce expression libraries by slicing cDNA inserts immediately adjacent to a bacterial promoter region on the vector; in these expression libraries, eukaryotic proteins are made in bacterial cells, which allows several important technological applications that are discussed below in DNA sequencing.

Several bacterial viruses have also been used as vectors. The most commonly used is lambda phage. The central part of the lambda genome is not essential for the virus to replicate in *E. coli,* so this can be excised using an appropriate restriction enzyme, and inserts from donor DNA are spliced into the gap. In fact, when the phage repackages DNA into its protein capsule, it includes only DNA fragments the same length of the normal phage genome.

Vectors are chosen depending on the total amount of DNA that must be included in a library. Cosmids are vectors that are hybrids of plasmid and phage and can carry larger inserts than either alone. Bacterial artificial chromosomes (BACs) are vectors based on F-factor (fertility factor) plasmids of *E. coli* and can carry much larger amounts of DNA. Yeast artificial chromosomes (YACs) are vectors based on autonomously replicating plasmids of *Sacchromyces cerevisieae* (baker's yeast). In yeast (a eukaryotic organism), a YAC behaves like a yeast chromosome and segregates properly into daughter cells. These vectors can carry the largest inserts of all and are used extensively in cloning large genomes such as the human genome.

Isolating the clone. In general, cloning is undertaken in order to obtain the clone of one particular gene or DNA sequence of interest. The next step after cloning, therefore, is to find and isolate that clone among other members of the library. If the library encompasses the whole genome of an organism, then somewhere within that library will be the desired clone. There are several ways of finding it, depending on the specific gene concerned. Most commonly, a cloned DNA segment that shows homology to the sought gene is used as a probe. For example, if a mouse gene has already been cloned, then that clone can be used to find the equivalent human clone from a human genomic library. Bacterial colonies constituting a library are grown in a collection of Petri dishes. Then a porous membrane is laid over the surface of each plate, and cells adhere to the membrane. The cells are ruptured, and DNA is separated into single strands—all on the membrane. The probe is also separated into single strands and labeled, often with radioactive phosphorus. A solution of the radioactive probe is then used to bathe the membrane. The single-stranded probe DNA will adhere only to the DNA of the clone that contains the equivalent gene. The membrane is dried and placed against a sheet of radiation-sensitive film, and somewhere on the film a black spot will appear, announcing the presence and location of the desired clone. The clone can then be retrieved from the Petri dishes.

DNA SEQUENCING

Once a segment of DNA has been cloned, its nucleotide sequence can be determined. The nucleotide sequence is the most fundamental level of knowledge of a gene or genome. It is the blueprint that contains the instructions for building an organism, and no understanding of genetic function or evolution could be complete without obtaining this information.

Uses. Knowledge of the sequence of a DNA segment has many uses, and some examples follow. First, it can be used to find genes. If a region of DNA has been sequenced, it can be screened for characteristic features of genes. For example, open reading frames (ORFs), long sequences that begin with a start codon and are uninterrupted by stop codons (except for one at its termination), suggest a protein-coding region. Also, human genes are generally adjacent to clusters of cytosine and guanine nucleotides (CpG islands). If a gene with a known phenotype (such as a disease gene in humans) is known to be in the chromosomal region sequenced, unassigned genes in the region will become candidates for that function. Second, homologous DNA sequences of different organisms can be compared to plot evolutionary relatedness both within and between species. Third, a gene sequence can be screened for functional regions. In order to determine the function of a gene, various domains can be identified that are common to proteins of similar function. For example, certain amino acid sequences within a gene are always found in proteins that span a cell membrane; such amino acid stretches are called transmembrane domains. If a transmembrane domain is found in a gene of unknown function, it suggests that the encoded protein is located in the cellular membrane. Other domains characterize DNA-binding proteins. Several public databases of DNA sequences are available for analysis by any interested individual.

Methods. The two basic sequencing approaches are the Maxam-Gilbert method and the Sanger method. In the most commonly used method, the Sanger method, DNA chains are synthesized on a template strand, but chain growth is stopped when one of four possible dideoxy nucleotides, which lack a 3′-hydroxyl group, is incorporated, preventing the addition of another nucleotide. A population of nested, truncated DNA molecules results that represents each of the sites of that particular nucleotide in the template DNA. These molecules are separated in a procedure called electrophoresis (Figure 14), and the inferred nucleotide sequence is deduced using a computer.

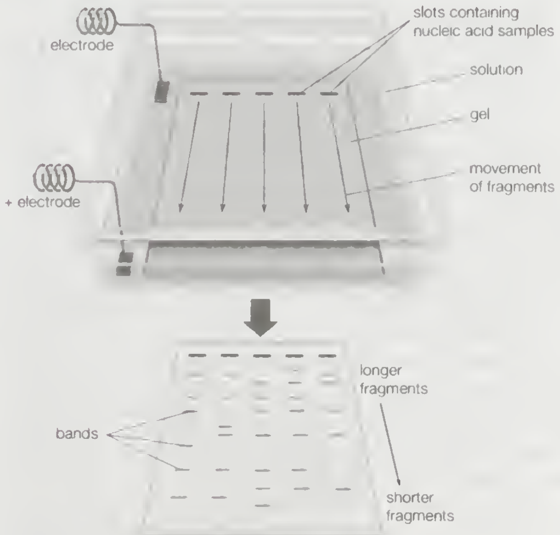


Figure 14: Electrophoresis is a technique that allows for the isolation and identification of nucleic acid samples.
Encyclopædia Britannica, Inc

In vitro mutagenesis. Another use of cloned DNA is in vitro mutagenesis. Mutations are useful to geneticists in enabling them to investigate the components of any biological process. However, traditional mutational analysis relied on the occurrence of random spontaneous mutations—a hit-or-miss method in which it was impossible to predict the precise type or position of the mutations obtained. In vitro mutagenesis, however, allows specific mutations to be tailored for type and for position within the gene. A cloned gene is treated in the test tube (in vitro) to obtain the specific mutation desired, and then this fragment is reintroduced into the living cell, where it replaces the resident gene.

One method of in vitro mutagenesis is oligonucleotide-directed mutagenesis. A specific point in a sequenced gene is pinpointed for mutation. An oligonucleotide, a short stretch of synthetic DNA of the desired sequence, is made chemically. For example, the oligonucleotide might have an adenine in one specific location instead of the wild-type guanine. This oligonucleotide is hybridized to the complementary strand of the cloned gene; it will hybridize despite the one base pair mismatch. Various enzymes are added to allow the oligonucleotide to prime the synthesis of a complete strand within the vector. When the vector is introduced into a bacterial cell and replicates, the mutated strand will act as a template for a complementary strand that will also be mutant, and thus a fully mutant molecule is obtained. This fully mutant cloned molecule is then reintroduced into the donor organism, and the mutant DNA replaces the resident gene.

Another version of in vitro mutagenesis is gene disruption, or gene knockout. Here, the resident functional gene is replaced by a completely nonfunctional copy. The advantage of this technique over random mutagenesis is that specific genes can be knocked out at will, leaving all other genes untouched by the mutagenic procedure.

Designer organisms

Transgenesis. The ability to obtain specific DNA clones using recombinant DNA technology has made it possible to add the DNA of one organism to the genome of another. The added gene is called a transgene. The transgene inserts itself into a chromosome and is passed to the progeny as a new component of the genome. The resulting organism carrying the transgene is called a transgenic organism or a genetically modified organism. In this way, a "designer organism" is made that contains some specific change required for an experiment in basic genetics or for improvement of some commercial strain. Several transgenic plants have been produced. Genes for toxins that kill insects have been introduced in several species, including corn and cotton. Bacterial genes that confer resistance to herbicides also have been introduced into crop plants. Other plant transgenes aim at improving the nutritional value of the plant.

Gene therapy. Gene therapy is the introduction of a normal gene into an individual's genome in order to repair a mutation that causes a genetic disease. When a normal gene is inserted into a mutant nucleus, it most likely will integrate into a chromosomal site different from the defective allele; although this may repair the mutation, a new mutation may result if the normal gene integrates into another functional gene. If the normal gene replaces the mutant allele, there is a chance that the transformed cells will proliferate and produce enough normal gene product for the entire body to be restored to the undiseased phenotype. So far, human gene therapy has been attempted only on somatic (body) cells. Somatic cells cured by gene therapy may reverse the symptoms of disease in the treated individual, but the modification is not passed on to the next generation. Germinal gene therapy aims to place corrected cells inside the germ line (*e.g.*, cells of the ovary or testis). If this is achieved, these cells will undergo meiosis and provide a normal gametic contribution to the next generation. Germinal gene therapy has been achieved experimentally in animals but not in humans.

Reverse genetics. Recombinant DNA technology has made possible a type of genetics called reverse genetics. Traditionally, genetic research starts with a mutant phenotype and, by Mendelian crossing analysis, a researcher is able to attribute the phenotype to a specific gene. Reverse genetics travels in precisely the opposite direction. Researchers begin with a gene of unknown function and use molecular analysis to determine its phenotype. One important tool in reverse genetics is gene knockout. By mutating the cloned gene of unknown function and using it to replace the resident copy or copies, the resultant mutant phenotype will show which biological function this gene normally controls.

Diagnostics. Recombinant DNA technology has led to powerful diagnostic procedures useful in both medicine and in forensics. In medicine, these diagnostic procedures are used in counseling prospective parents as to the likelihood of having a child with a particular disease, and they are also used in the prenatal prediction of genetic disease in the fetus. Researchers look for specific DNA fragments that are located in close proximity to the gene that causes the disease of concern. These fragments, called restriction fragment length polymorphisms (RFLPs) often serve as effective "genetic markers." In forensics, DNA fragments called variable number tandem repeats (VNTRs), which are highly variable between individuals, are employed to produce what is called a "DNA fingerprint." A DNA fingerprint can be used to determine if blood or semen left at the scene of a crime belongs to a suspect.

DNA fingerprinting

Genomics. The genetic analysis of entire genomes is called genomics. Such a broad-scale analysis has been made possible by the development of recombinant DNA technology. In humans, knowledge of the entire genome sequence has facilitated searching for genes that produce hereditary diseases. It is also capable of revealing a set of proteins, produced at specific times, in specific tissues, or in specific diseases, that might be targets for therapeutic drugs. Genomics also allows the comparison of one genome with another, leading to insights into possible evolutionary relationships between organisms.

Genomics has two subdivisions, structural genomics and functional genomics. Structural genomics is based on the complete nucleotide sequence of a genome. Each member of a library of clones is physically manipulated by robots and sequenced by automatic sequencing machines, enabling a very high throughput of DNA. The resulting sequences are then assembled by a computer into a complete sequence for every chromosome. The complete DNA sequence is scanned by computer to find the positions of ORFs, or prospective genes. The sequences are then compared to the sequences of known genes from other organisms, and possible functions are assigned. Some ORFs remain unassigned, awaiting further research.

Functional genomics attempts to understand function at the broadest level (the genomic level). In one approach, gene functions of as many ORFs as possible are assigned as above in an attempt to obtain a full set of proteins encoded by the genome (called a proteome). The proteome broadly defines all the cellular functions used by the organism. Function in relation to specific developmental stages also is assessed by trying to identify the "transcriptome," the set of mRNA transcripts made at specific developmental stages. The practical approach ultilizes microarrays, glass plates the size of a microscope slide imprinted with tens of thousands of ordered DNA samples, each representing one gene (either a clone or a synthesized segment). The mRNA preparation under test is labeled with a fluorescent dye, and the microarray is bathed in this mRNA. Fluorescent spots appear on the array indicating which mRNAs were present, thus defining the transcriptome.

Protein manufacture. Recombinant DNA procedures have been used to convert bacteria into "factories" for the synthesis of foreign proteins. This technique is useful not only for preparing large amounts of protein for basic research but also for producing valuable proteins for medical use. For example, the genes for human proteins such as growth hormone, insulin, and blood-clotting factor can be commercially manufactured. Another approach to producing proteins via recombinant DNA technology is to introduce the desired gene into the genome of an animal, engineered in such a way that the protein is secreted in the animal's milk, facilitating harvesting.

(A.J.F.G.)

HEREDITY AND EVOLUTION

The gene in populations

Influences on heredity

In the study of heredity the first question that arises is how the genotype of an individual is formed from the constituents of the genotypes of his parents. This is the genetics of individuals or basic genetics. One may also inquire how the genotype in a fertilized egg cell influences the developmental pattern of the organism and thus realizes its potentialities. This is developmental genetics. An individual, at least an individual of a sexually reproducing species, is not, however, biologically complete in itself. Its biological role is actualized through its membership in a reproductive community, a Mendelian population. A Mendelian population consists of individuals among whom matings may or do occur. An individual is mortal and temporary; a Mendelian population has a continuity through time. The genetic processes in Mendelian populations are the subject matter of population genetics.

THE GENE POOL

A Mendelian population is said to have a gene pool. The gene pool is the sum total of the genes carried by the individual members of the population. The gene pool also continues through time. The genes of the individuals of the generation now living come from a sample of the genes of the previous generation; if these individuals reproduce, their genes will pass into the gene pool of the following generations. The Mendelian population and its gene pool in humans have a very complex structure. Individuals born and living close together are more likely to meet and to mate than those living far apart. In a widely distributed species such as *Homo sapiens,* the likelihood of mating of individuals born on different continents was, until the development of modern means of travel, very small. The gene pool of the human species is, accordingly, divided into the smaller gene pools of populations living in different regions. Aside from the geographic divisions, there are also linguistic, religious, social, economic, and educational barriers that break the gene pools into further, often overlapping, subdivisions. The smallest subdivision is referred to as an isolate or panmictic unit; it consists of a relatively limited number of persons (or animals or plants) that may be regarded as potential mates. Few of these divisions may be sharp enough to decide where one gene pool subdivision ends and the other begins, and yet these subdivisions are biologically meaningful.

A biological species, in sexually reproducing organisms, is defined as the most inclusive Mendelian population. The gene pool of *Homo sapiens* is an entity the limits of which are not in doubt, since no gene exchange between the human and any other related species takes place. Nor does the intraspecific differentiation impair the unity. There may never have been a marriage of, for example, an Eskimo and a Melanesian, but genetic communications between the Eskimo gene pool and the Melanesian gene pool occur through the chains of geographically intermediate populations. A genetic change arising anywhere in the world, if favourable, may spread throughout humanity. This is how genetic changes may have transformed the ancestral prehuman species into the present one. This genetic unity makes any genetic damage (*e.g.,* that caused by exposure to high-energy radiation) a concern of all people, regardless of whether the damage is inflicted more heavily on one portion of the human population than on another.

THE HARDY–WEINBERG PRINCIPLE

The basis for population genetics

In 1908, Godfrey Harold Hardy and Wilhelm Weinberg independently formulated a theorem that became the foundation of population genetics. According to the Hardy–Weinberg principle, two or more gene alleles will have the same frequency in the gene pool generation after generation, until some agent acts to change that frequency. Consider a population that is, as most human populations actually are, a mixture of individuals with M, N, and MN blood types. An individual witn M blood is a homozygote with two *M* alleles (*MM*), an N individual has two *N* alleles (*NN*), and an MN individual is a heterozygote (*MN*). Suppose that a population consists of 49 percent of individuals with M, 42 percent with MN, and 9 percent with N bloods. The frequencies of the blood types in the next and the following generations can be calculated. Assume for simplicity that (1) marriages are at random with respect to the blood groups, (2) people with different blood groups have neither advantages nor disadvantages in survival or in reproduction, (3) the alleles *M* and *N* do not change frequently by mutation, (4) there is no migration into or out of the population, and (5) the population is large enough so that chance fluctuations may be ignored. Also assume that all individuals produce equal numbers of sex cells with each of the pair of alleles they carry and that the sex cells of the parents combine at random in fertilization.

Persons with M blood produce sex cells with the allele *M,* and these sex cells will amount to 49 percent of the total. Persons with MN blood produce equal numbers of *M*- and of *N*-bearing sex cells—*i.e.,* 21 percent of each. Finally, persons with N blood will give *N*-bearing sex cells, 9 percent of the total. The gene pool will, therefore, contain $49 + 21 = 70$ percent of *M*- and $21 + 9 = 30$ percent of *N*-type sex cells, or, using decimals, 0.7 *M* and 0.3 *N,* respectively. These sex cells will combine to produce the following blood types in individuals: $0.7 \times 0.7 = 0.49$, or 49 percent of M; 0.3×0.3, or 9 percent of N; and $2 \times 0.7 \times 0.3$, or 42 percent of MN. Generalizing, if the proportions of *M* and *N* genes in the gene pool are p and q, respectively, the frequencies of the blood groups will be, generation after generation:

$$p^2 + 2pq + q^2 = 1.$$

(In this expression p^2 represents the genotype *MM,* q^2 the genotype *NN,* and pq the genotype *MN.*) This expression is the Hardy–Weinberg formula, which describes the genetic equilibrium status in populations. The genetic composition of a population can meaningfully be described in terms of the frequencies of various alleles of the genes in the gene pool. Different populations of the same species are likely to differ in the frequencies of some, probably of many, genes. If the gene frequencies are different, the populations are distinct; if the differences are large, one may decide to give these populations different subspecific names.

CHANGES IN GENE FREQUENCIES

The Hardy–Weinberg principle predicts that gene frequencies will remain constant from generation to generation within a population that meets certain assumptions. It further predicts that if mating is random in regard to genetic traits, then the frequencies of genotypes will also remain constant in succeeding generations. Yet if all gene frequencies remained constant in populations indefinitely, evolution could not take place. Evolution is, in the last analysis, change of gene frequencies.

Evolution

The assumptions that underlie the Hardy–Weinberg principle are, in fact, theoretical considerations that are never met in natural populations. Mutations and migrations can affect gene frequencies. Many natural populations are small enough that chance fluctuations can significantly alter gene frequencies. Moreover, in many cases matings may be selective rather than random in regard to certain genetic traits; this will alter the frequencies of genotypes in succeeding generations. The effects of these phenomena are described in this section.

Finally, and most importantly in considering the genetics of evolution, a certain genotype may offer an advantage in terms of reproduction or survival over its counterparts. Such advantageous genotypes, and their constituent alleles, will increase in frequency in succeeding generations. This phenomenon, called selection, is dealt with at length below (see *Selection as an agent of change*).

Mutations. As has been discussed, under most circumstances newly formed chromosomes and their genes are perfect reproductions of the originals. This remarkably stable copying process is the basis of the continuity of living species, generation after generation, for each specific characteristic encoded by the genes.

It also has been shown, however, that mutations—in the form of chromosomal aberrations and miscopying of DNA—do occur. Mutations produce the potential for new traits. Even when only a single gene in one cell is mutated, the copying process tends faithfully to reproduce the changed DNA in all the descendants of that particular cell. Indeed, mutation appears to be a basic cellular mechanism underlying evolution.

That mutations can arise in response to environmental agents was first demonstrated by Hermann J. Muller, who in 1926 demonstrated that *Drosophila* flies exposed to X rays suffer high rates of mutation. It is now known that other forms of radiation, as well as a variety of chemicals, can serve as potent mutagenic agents, producing both chromosomal aberrations and changes in the DNA of individual genes. The implications of these findings are discussed below in *Human genetics.*

In discussing mutation in the context of gene frequencies and evolution, it is imperative to recall the difference between germinal mutations and somatic mutations. Only the former are passed on to offspring, who will then "breed true" for the altered or defective trait. It is germinal mutations, then, that can produce the greatest effect in altering gene frequencies in succeeding generations of a population.

Gene flow. Migrations of individuals into a population can introduce new alleles or can increase the frequencies of those already present; similarly, migration out of a population can remove alleles or decrease their frequencies. This gene flow may be negligible in a highly isolated population, such as one on a remote island, but it operates freely among adjacent populations of species that occupy large ranges. Gene flow is most readily appreciated in human and other animal populations made up of mobile individuals; it occurs in plants as well, however, as pollen may be carried by wind or by animals from one population to another.

Genetic drift. This term refers to the effects of chance fluctuations on gene frequencies in small populations. Consider a small population in which there are two alleles—*a* and *A*—for a particular gene. If the frequency of one of these alleles, say *a,* is low to begin with, a chance event that has nothing to do with the selective value of that allele could result in its complete elimination from the population. The population would then consist entirely of *AA* individuals. Such an event is referred to as a fixation for gene *A*. Conversely, a chance event could result in the loss of several *AA* individuals from the population; in this case, the frequency of *a* would increase.

Founder effect

Genetic, or random, drift is most obviously manifested in what is known as the "founder effect." This occurs when a small group of individuals—or even just one pregnant female—migrates into a new region, or when a small group becomes reproductively isolated from its parent population without migration. The genes carried by the founders of the new population will often be small, atypical, and unbalanced samples of the gene pool of the population whence they came. In the case of a new, small population established by migration, the subsequent frequencies of alleles will depend not only on the chance distributions carried by the founders but also on the environmental influences (*i.e.,* the selection pressures) of the new location. In the case of reproductive isolation without migration, the situation depends more on the chance distribution of alleles in the small group at the time of isolation and the relative degree of consanguinity among those mating. These factors might explain the increased frequency of a relatively rare autosomal recessively inherited disease, Tay-Sachs, among Jews of eastern European origin. For many years these people were forced to live in small isolates and were socially ostracized by the population around them. As a result, a mutant gene present in the original small group had the opportunity to express itself as a lethal disease among homozygotes.

Selective mating. Mating may be selective rather than random with respect to a given gene. Suppose that persons with M, MN, and N bloods prefer to marry individuals with a blood group the same as themselves. Selective mating will not, by itself, change the gene frequencies, but it will disturb the Hardy–Weinberg equilibrium in another way. The relative frequencies of the homozygotes (*MM* and *NN*) will increase from generation to generation, while those of the heterozygotes (*MN*) will decrease. Eventually the population will consist of homozygotes only. Preferential mating of unlike genotypes will, on the contrary, increase the incidence of the heterozygotes, but it will not, no matter how long continued, eliminate the homozygotes.

Selection as an agent of change

NATURAL SELECTION AND DARWINIAN FITNESS

Sexual reproduction under simple (Mendelian) inheritance is a conservative force that tends to maintain the genetic status quo in a population. If a gene frequency is 1 percent in a population, it tends to remain at 1 percent indefinitely unless some force acts to change it. Outside of the laboratory, the most powerful force for changing gene frequencies is natural selection.

The carriers of some genes may survive more often or be more fecund than the carriers of other genes. When the carriers of different genes are not equally efficient in transmitting these genes to the succeeding generations, the result is natural selection. When the inequality of the transmission rates of the genes is imposed by human will, the result is artificial selection. In general, the genes that confer on their possessors a superior reproductive efficiency will increase in frequencies from generation to generation, and the reproductively inferior genes will become less frequent.

Imagine a population that carries two alleles—A_1 and A_2—for a particular gene. Suppose that the relative numbers of the surviving progeny left by the carriers of the genotypes A_1A_1 and A_2A_2 are in the ratio $1 : 1 - s$ (the value s is called the selection coefficient). If for every 100 offspring of A_1A_1 parents only 90 surviving offspring are left by A_2A_2 parents, then $s = 1/10$ or 0.1. The heterozygotes, A_1A_2, may leave as many progeny as A_1A_1 (if A_1 is dominant) or as many as A_2A_2 (if A_2 is dominant) or an intermediate number (if neither is dominant). Alternatively, the heterozygotes may exhibit the quality of hybrid vigour (heterosis); that is, they may be reproductively superior to both homozygotes A_1A_1 and A_2A_2. And finally (though this is rare except in species hybrids) the heterozygotes may be at a disadvantage compared to both homozygotes. The situation is simplest if the heterozygotes (A_1A_2) are equal in reproductive efficiency to one of the homozygotes or intermediate between the two.

Hybrid vigour

Whichever gene, A_1 or A_2, confers a superior reproductive efficiency on its possessors will increase in frequency in the population. The increase will continue generation after generation; given enough time (*i.e.,* enough generations) the more efficient gene will eliminate and supplant the less efficient one entirely. How rapid or slow the gene frequency changes will be depends on the magnitude of the selection coefficients. Table 4 gives several examples for a dominant allele, for a recessive allele, and for no dominance—*i.e.,* for the case in which the fitness of the heterozygote is intermediate between the two homozygotes. The two alleles in the population with which the selection starts are assumed to be equally frequent, $p = q = 0.5$. The selection coefficients of one (lethal), 0.5, 0.1, and 0.01 are considered, and the gene frequencies after one, two, five, 10, 20, 50, and 100 generations of such selection are given.

The homozygotes for a recessive gene allele may not reproduce at all. With natural selection this may happen because they are inviable (lethal) or sterile; with artificial selection the same result is accomplished if the breeder kills them or does not use them as parents. The recessive allele is then opposed by a selection $s = 1$. Table 4 shows that a gene with an initial frequency of 0.5 will decrease to 0.33 after one generation, to 0.25 after two, and to 0.01 after 100 generations. With weaker selection (s smaller than unity) the decrease of the recessive allele will, of course, be slower. Note, however, that in all cases the frequency change is more rapid when the gene is common than when

Table 4: The Decrease of the Frequencies of Genes Discriminated Against by Different Selection Pressures
(the frequency of the gene in the initial population is 0.5)

generations	strong selection	intermediate selection		weak selection
	(s = 1.0)	(s = 0.5)	(s = 0.1)	(s = 0.01)
		recessive allele		
1	0.33	0.43	0.49	0.499
2	0.25	0.37	0.48	0.497
5	0.14	0.27	0.44	0.494
10	0.08	0.16	0.39	0.488
20	0.05	0.09	0.31	0.476
50	0.02	0.04	0.18	0.442
100	0.01	0.02	0.10	0.391
		dominant allele		
1	—	0.40	0.49	0.499
2	—	0.29	0.47	0.497
5	—	0.07	0.43	0.494
10	—	0.002	0.35	0.487
20	—	—	0.20	4.474
50	—	—	0.01	0.433
100	—	—	—	0.362
		intermediate heterozygote		
1	0.25	0.42	0.49	0.499
2	0.13	0.34	0.47	0.497
5	0.02	0.17	0.43	0.494
10	0.001	0.04	0.37	0.487
20	—	0.003	0.26	0.475
50	—	—	0.07	0.438
100	—	—	0.006	0.377

it is rare. A dominant allele opposed by a selection s = 1 (a dominant lethal) disappears in a single generation, and even weaker selections against dominants are more efficient than similar selections against recessives. A selection in favour of a recessive is, of course, just the reverse of that against a dominant, and vice versa. The frequencies can be read from Table 4 by subtracting the frequencies given from unity. When the dominance is absent, the efficiency of selection is, as shown in Table 4, intermediate between those for recessives and for dominants.

The attainment of balanced polymorphism

A most interesting, and at first sight paradoxical, outcome of selection arises if the heterozygote is superior to both homozygotes. Neither the gene A_1 nor A_2 is allowed to crowd the other out or to disappear entirely. Instead, a genetic equilibrium is reached, and the population attains the state of the so-called balanced polymorphism. All three genotypes continue to occur in the population, with frequencies dependent on the relative magnitudes of the selection coefficients s_1 and s_2. This will be true even if one of the homozygotes is seriously incapacitated, inviable, or sterile. The possible importance of this in human populations is considered below.

Darwin's description of the process of natural selection as the survival of the fittest in the struggle for life is a metaphor. "Struggle" does not necessarily mean contention, strife, or combat; "survival" does not mean that ravages of death are needed to make the selection effective; and "fittest" is virtually never a single optimal genotype but rather an array of genotypes that collectively enhance population survival rather than extinction. All these considerations are most apposite to consideration of natural selection in humans. Decreasing infant and childhood mortality rates do not necessarily mean that natural selection in the human species no longer operates. Theoretically, natural selection could be very effective if all the children born reached maturity. Two conditions are needed to make this theoretical possibility realized: first, variation in the number of children per family. and, second, variation correlated with the genetic properties of the parents. Neither of these conditions is far-fetched.

Darwinian fitness is sometimes referred to also as the adaptive value or the selective value; these terms are best treated as synonyms, although they may have somewhat different connotations. The Darwinian fitness of a genotype, or of a group of genotypes, is measured as the contribution of their carriers to the gene pool of the succeeding generation, relative to the contributions of other genotypes present in the same population. In the example given above, the Darwinian fitness of the genotype A_1A_1 was taken to be unity and the fitness of A_2A_2 as 1 − s or less than unity. The fitness is, of course, subject to change in different environments; the carriers of a genotype A_1A_1 may leave more surviving progeny than A_2A_2 in a certain environment, but the reverse may be the case in another environment. Darwinian fitness is reproductive fitness; bodily or mental vigour, health, and energy obviously contribute to this fitness but only insofar as they result in a superior reproductive capacity. Mules, no matter how strong and resistant, must be ranked zero in Darwinian fitness because they are sterile. The emphasis on reproductive success rather than on survival is characteristic of the modern concept of natural selection, as distinguished from the classical one. The difference is not, however, so great as it may seem at first glance; the carriers of a genotype evidently must survive in order to reproduce, and they must reproduce in order to survive in the next generation.

VARIETIES OF NATURAL SELECTION

Normalizing selection

There are several kinds of natural selection, rather different in their biological consequences and in their importance to humans. The simplest of them is the normalizing selection, which was already known before Darwin but, of course, not under this name. Normalizing selection counteracts the accumulation in populations of hereditary diseases, malformations, and weaknesses. Suppose that a gene allele A_1, the carriers of which have a high Darwinian fitness, mutates to a state A_2, which lowers the fitness. If A_2 is a dominant lethal or a gene that renders its carriers sterile, then (as shown in Table 4, column s = 1.0) all the A_2 mutants will be eliminated in the same generation in which they arise. A new crop of mutants will, of course, appear in the next generation. If, however, the selection is not so completely efficient, some mutant genes will escape its dragnet and will be transmitted to the next generation. That generation will contain all the newly arisen mutants, a part of the mutants that arose in the preceding generation, a smaller part of those having arisen two generations ago, etc. How great a "genetic load" of uneliminated mutants a population can accumulate will depend principally on two factors—how often the mutation arises and how much it lowers the Darwinian fitness.

Simple formulas have been worked out to describe the situations that arise. Suppose that a deleterious mutation from $A_1 \rightarrow A_2$ occurs at a rate u per generation. Suppose further that the mutant is discriminated against by a selection coefficient s. If, then, the mutant A_2 is dominant to the original state, A_1, the frequency of A_2 in the gene pool will be u/s. If A_2 is recessive to A_1, its accumulated frequency will be much higher, namely $\sqrt{u/s}$. The reason deleterious recessive mutants are allowed to attain higher frequencies than equally deleterious dominants is simple: a recessive mutant may be carried in many heterozygotes, in which it does not express itself, and is consequently protected, or sheltered, from the weeding-out action of natural selection. With mutants that are neither dominant nor recessive, the accumulation will be intermediate between u/s and $\sqrt{u/s}$.

All human populations doubtless carry genetic loads consisting of harmful mutant genes. This cannot be blamed entirely on culture, civilization, or on any other specifically human attributes. Populations of *Drosophila* flies and of other sexually reproducing organisms also carry genetic loads. The accumulation of the genetic loads is a necessary consequence of the occurrence of mutations, most of which are harmful but not always harmful enough to be eliminated immediately after they are produced. Harmful mutations are accumulated until the numbers of the respective mutant genes become equal to the numbers eliminated by natural selection in the same population. The population is then said to be in the state of "genetic equilibrium." Muller has termed the elimination of harmful mutants "genetic death." Genetic "death" is sometimes cruel, sometimes rather benign. The death of a child from a severe hereditary disease and the genetically conditioned failure to have one more child are both genetic deaths. The higher the mutation rates, the more harmful are the mutants produced and the more frequent are the genetic deaths. In populations that have reached genetic equilibrium, the total number of genetic deaths will be equal to the total number of the mutations subject to normalizing natural selection.

Heterotic balancing selection

A very different form of natural selection is heterotic balancing selection. It occurs when the Darwinian fitness of a heterozygote exceeds the fitness of both homozygotes, a situation mentioned above. Heterotic balancing selection also leads to genetic equilibrium but not to an equilibrium between mutation and the normalizing selection. The balanced polymorphism that is established is due to the selection favouring the heterozygotes against the homozygotes. In a sexually reproducing population the heterozygotes tend, however, to produce a fresh crop of homozygotes in every generation. The maintenance in human populations of the grave hereditary disease sickle-cell anemia is apparently due to this form of selection. The sickling gene (Hb^S) produces a specific type of hemoglobin, while normal hemoglobin is related to another allele (Hb^A). Accordingly, the possible genotypes are Hb^AHb^A, Hb^AHb^S, and Hb^SHb^S. The latter individuals are homozygous for the sickle-cell gene and will develop severe anemia. While the condition is not lethal before birth, such individuals rarely survive long enough to exhibit more than minimal fitness in the Darwinian sense of capacity to reproduce. On these grounds it might be concluded that natural selection eventually should drive the frequency of the Hb^S gene to complete elimination or at least down to the one in 100,000 level of the mutation rate. One would theorize a transient polymorphism; that is, one on the way out, toward fixation of the favoured allele.

Evidence, however, seems to contradict theory since, in a number of African tribes living in their ancestral tropical lowlands where the falciparum form of malaria is widespread, the Hb^S (sickling) gene is very common indeed. On the other hand, the same gene is rare in genetically related but isolated tribes living in highlands that are free of mosquitoes that transmit this type of malaria. This discrepancy between lowland and highland people has led to the hypothesis that the Hb^AHb^S heterozygote is fitter and capable of leaving more offspring than is the homozygous normal Hb^AHb^A in a highly malarious environment. This extra measure of protection is evidently provided by the sickle-cell hemoglobin (Hb^S), which is detrimental to the malaria parasite. In malarial environments, populations that contain the sickle-cell gene, therefore, have advantages over populations free of this gene. The former populations are in less danger from the ravages of malaria, although they "pay" for this advantage by sacrificing in every generation some individuals who die of anemia. The lethal disease caused by homozygosity for the sickle-cell gene certainly brings about some genetic deaths; it is a part of the genetic load of the populations. But this genetic load, due to a disadvantage of being homozygous for certain genes, is very different from the mutational load controlled by the normalizing selection. The former is maintained by the heterotic balancing selection, while the latter is maintained by recurrent mutation.

Diversifying selection

Another form of natural selection is diversifying, or disruptive, selection. In many discussions and mathematical analyses of selection this simplifying assumption is adopted: that the environment in which a population lives is uniform and that the selection advantages and disadvantages of different genotypes are independent of their frequencies in the population. This simplification, however, flies in the face of reality. Many animals can subsist on a variety of foods; many plants grow on different soils; humans have to fill many different employments, functions, professions, and social roles. It is most likely that some genotypes will be fitter in some environments than they are in other environments. Diversifying selection will then favour different genotypes in different subenvironments, or ecological niches, that occur in the population.

A special form of selection occurs in mammals due to the incompatibility of certain maternal genotypes with those of their unborn children. The best studied case in humans is that of a Rhesus-positive fetus in a Rhesus-negative mother (see below *The genetics of human blood*). This selection should, theoretically, make the entire population either Rhesus positive or Rhesus negative. For reasons that have not been clarified, it does not appear to be doing this. Another special kind of selection is that due to so-called meiotic drives, disturbances of the Mendelian segregation mechanisms, which result in sex cells carrying certain gene alleles being more or less frequent than expected on a random basis.

Directional selection

The last to be mentioned, but in the long run possibly the most important form of selection, is directional selection. Suppose that the climate becomes warmer or cooler, that there appears a new source of food or a new predator or disease, or that there occur some other prominent environmental changes. Some genotypes will, then, become more favourable and others less favourable. Directional natural selection will operate to reconstruct the gene pool of the population in accord with the demands of the new environment.

NATURAL SELECTION IN OPERATION

When the antibiotic streptomycin is added in a sufficient concentration to a culture of colon bacteria, only the streptomycin-resistant mutants are able to reproduce, while the streptomycin-sensitive cells are eliminated. If experimenters add the streptomycin to a culture in order to obtain resistant strains of the bacteria, they are effecting artificial selection. Natural selection acts, however, very much in the same fashion. Widespread and often rather indiscriminate utilization of antibiotics against various infectious microorganisms led to quite unintended consequences—the appearance of numerous "drug-fast" infections. In some instances (as with gonorrhea in some countries) the treatment with penicillin, originally very effectual, became less reliable or even useless as selection brought about a drug-resistant strain of gonococci.

Genetic resistance to DDT

Essentially the same process took place with more complex organisms, various insect pests, in response to the introduction of certain potent insecticides. The first such insecticide, DDT, was used on a large scale during and after World War II. The discovery of DDT, followed by invention of many other insecticides of similar or even greater potency, led to hope of eradication of all insect pests. This enthusiasm was short lived, since by 1947 DDT-resistant populations of the housefly were recorded in Italy and soon thereafter in other countries. It apparently takes the fly only two to three years to evolve a genetically conditioned resistance to DDT, and similar resistances arise against other insecticides as well. During the Korean War DDT-resistant lice emerged. Attempts to eradicate malaria-bearing mosquitoes by DDT sprays resulted in DDT-resistant mosquito populations. Laboratory experiments with *Drosophila* and with houseflies have shown that one can artificially select for insecticide resistance. Moreover, there are apparently genetically different resistant strains, some differing from the sensitive ones by a single gene with a strong effect and others by accumulation of several genetic changes, each increasing the resistance only slightly but giving strong resistance in the aggregate.

Industrial melanism

Colour changes in moth populations provide another example of natural selection. A black mutant of the peppered moth (*Biston betularia*) was found at Manchester, England, in 1848. Until that time the prevalent form of this species was light gray with dark speckles. The frequency of the black variant rapidly increased in the immediate area and reached about 99 percent by 1898. Dark coloration, melanism, has arisen and spread in many species of British moths belonging to different genera and families. The spread of the melanic forms was most rapid in and near industrial areas, where the foliage and the tree trunks are blackened with soot and other atmospheric pollutants. The spread of the dark forms is consequently known as industrial melanism. The light form of *Biston betularia* is protectively coloured when resting on tree trunks covered with lichens but conspicuous on blackened trunks on which the lichens have been largely destroyed by pollution. The selective agents are the insectivorous birds that feed on the moths; they find the light moths easily on dark tree bark and the black moths easily on light tree trunks. H.B.D. Kettlewell has verified the hypothesis by exposing known numbers of light and of melanic moths in localities with blackened and with unpolluted vegetation. A greater proportion of the blacks than of the whites remained alive and were recaptured on dark vegetation, and the opposite was observed on the unpolluted vegetation. Genetically, the difference between the light and the melanic forms has been shown to be due to a single gene, the allele for melanism being dominant to that for the light coloration.

Plants that grow in habitats created or modified by human activities are often different from their wild relatives and different in ways that adapt them to survive and reproduce in their respective habitats. *Camelina sativa,* a plant of the mustard family, is a common weed in fields of cultivated flax in Europe. The *Camelina* that contaminates the flax fields is, however, genetically different in several ways from that growing outside the fields. *Camelina* outside flax fields is a low-growing, branched plant with small seeds; that growing in flax fields is much taller, unbranched, and with larger seeds that resemble those of the flax in size and in specific gravity. The characteristics of *Camelina* growing in flax fields are such that it is harvested together with the flax, and its seeds remain with those of flax during the winnowing process. The variety in the flax fields is doubtless derived from the original form now found growing freely; the remarkable adaptations that are found in the former have, then, developed since farmers started to plant flax on fairly large scales. Genetically, the varieties of *Camelina* differ from each other in several, probably in many, genes. The process of natural selection—incidentally abetted by the human activity of flax cultivation—has, thus, created a new adaptive genetic system.

The examples of drug resistance in bacteria, pesticide resistance in insects, and industrial melanism in moths all involve directional natural selection, which, if continued long enough, leads to replacement of old gene alleles by new ones. Balancing natural selection has a different biological function—it maintains genetic diversity, or polymorphism, in populations. Selection pressures involved in balancing selection in nature are sometimes quite high. This is very useful to those engaged in research on selection, since it is evidently easier to detect strong selection than weak selection.

GENETICS OF SPECIES DIFFERENCES

The nature and origin of species are dealt with in the article EVOLUTION, THE THEORY OF, particularly in the section *Species and speciation.* Here it is necessary to consider only the genetic composition of species differences in sexually reproducing and outbreeding organisms. In general, species are considered to be populations of organisms between which breeding is impossible or significantly limited under natural conditions. (This limitation does not hold for the mating [hybridization] of cells of different species by the methods of somatic cell genetics [see above].

An individual belongs to only one species, unless that individual is a species hybrid. Mules, hybrids between the horse and donkey, are sterile because of abnormalities in the processes of sex-cell formation in their gonads. Sterility of hybrids between species, if viable hybrids between them can be obtained at all, is observed very frequently, though some experimentally obtained species hybrids have proved to be fertile.

The origin of distinct populations

Scientists have asked what causes the development of the gene-frequency differences between populations that live in different territories, or, in other words, what makes these populations distinct. The probable explanation is that genetic differences between populations arise in most cases through natural selection in response to the local environments that prevail in the territories they inhabit. It is, however, very difficult to verify this explanation in many concrete instances of race differentiation. For example, it is probable that the dark skin pigmentation of many human populations that live, or have until recently lived, in tropical and subtropical countries protects them from sunburns. It is probable also that the light skin colours of the natives of Europe facilitate the acquisition of vitamin D in regions with deficient sunshine. The evidence for even these hypotheses is not as conclusive as might be desired. But when it comes to human traits such as hair form and shapes of the nose, lips, and cheekbones, no acceptable evidence of adaptive significance is available. The situation is no better with animals and plants: for most physical differences the adaptive significance is unknown.

Attempts have been made to envisage factors other than natural selection that could be responsible for genetic differentiation of populations. Appeals are frequently made to pleiotropism of the gene action; a visible difference may in itself be neither adaptive nor unadaptive, yet it may be only an outward sign of an underlying physiological difference that is adaptively important. An elegant example is the coloration of onions—red and purple bulbs are resistant to the attacks of a smudge fungus, while white bulbs are highly susceptible. A trait may also be important as a sexual recognition mark, or it may play a role in the courtship ritual.

A most interesting possibility that should be seriously considered is that some differences between populations may be due to random genetic drift. As was discussed above, genetic drift acts on small populations. Suppose that a species lives in many isolated colonies, some of them consisting of only tens or perhaps hundreds of individuals. Chance events may cause the gene frequencies in the different colonies to drift apart. How important this random genetic drift may be in race differentiation is controversial. That genetic drift does occur is certain; a simple example is that in small villages a sizable fraction of the inhabitants sometimes have the same surname, and different surnames are frequent in different villages. Increasing or decreasing frequencies of the surnames evidently go together with increases or decreases of the frequencies of certain genes that the ancestors of the people with these surnames carried. As discussed, genetic drift can also arise from the founder effect. When the populations of new colonies founded by a small number of individuals expand, they will be found to differ genetically from each other and from the ancestral population. Natural selection will then come into operation, giving rise to new balanced gene pools. The founder effect was probably important in the development of some human populations. Many groups may be the descendants of small numbers of original migrants and settlers. Whether the random genetic drift alone can explain the origin of the gene complexes that differentiate species is very doubtful. The point is, however, that genetic drift and natural selection are not mutually exclusive alternatives; it is not one or the other but the interaction of both that brings about differentiation. The founder effect is a special case of random genetic drift. The gene pool of a colony derived from a single immigrant or several pairs of immigrants may need a restructuring by natural selection to become properly adapted to the new environment.

OUTBREEDING AND INBREEDING

Human beings make up an outbreeding species. Marriage between close relatives is forbidden by custom and by law in most human societies. Whether the proscription initially was based on biological or social considerations is a matter of question. The prohibition may seem reasonable, however, in view of the evidence accumulated by practical breeders and geneticists. In normally outbreeding species, the progeny of matings in which the parents are close relatives tend to be less vigorous than the offspring of unrelated parents. This is a consequence of the genetic loads of deleterious recessive genes carried in populations. A recessive gene is more or less innocuous in the heterozygous condition, but its chances of becoming homozygous are greater in the offspring of parents who are close relatives (and therefore have more similar genotypes) than in families where the parents are not closely related.

The effects of outbreeding and inbreeding

It is worthwhile to stress at this point that the effects of outbreeding and inbreeding on the vigour, viability, and fertility of the offspring depend upon the reproductive biology of the species or the population in which they occur. Outbreeding is not invariably or necessarily invigorating, nor is inbreeding invariably or necessarily debilitating. Wheat is an example of a plant that is predominantly self-pollinating. Self-pollination is, of course, the closest possible form of inbreeding, but this inbreeding does not progressively weaken the vigour of a wheat strain. The converse, however, is observed with corn. Inbreeding decreases the yield, and the intercrossing of inbred lines restores hybrid vigour in the progeny. One may say that hybrid vigour (heterosis) is the normal state of a corn population, while inbreeding and a prevalence of homozygosis is the normal state in wheat. In species and populations in which the reproductive biology is adjusted to outbreeding, consanguinity leads to a decline in the average vigour and to the appearance of relatively many individuals with hereditary diseases and malformations. Just what kind of a malformation or ill health, if any, will appear in a given progeny depends on what deleterious recessive genes were carried by the common ancestor of the relatives who mate. This will certainly be different in different families. The histories of the reigning houses of Egypt and the Peruvian Incas, where the monarchs allegedly married their sisters, are not sufficiently well documented to conclude that in these instances the incestuous marriages resulted in no weakening of the progeny. High childhood mortality may have gone unrecorded; furthermore, "sister" may have been, in some instances, a title bestowed upon a person rather than indicative of an actual relationship.

It is important to study how much inbreeding depression will occur in a given species or population, and a great deal of information in this direction is being accumulated. With rare exceptions, the highest degree of consanguinity that occurs in human populations with appreciable frequency is marriage of first cousins, resulting in homozygosis of 6.25 percent of the genes in the progeny. A genetic counselor would hardly be justified in discouraging all marriages of people known to be relatives, and yet he may point out that the chance of genetic weaknesses in the progeny of such marriages is measurably greater than when unrelated persons marry.

GENETIC LOAD AND HYBRID VIGOUR

How hereditary variability is maintained in populations of sexually reproducing species is an unsettled and controversial question. The classical theory assumes that the pressure of mutation generating detrimental genetic variants is the main factor. The balance theory, while not denying that some variability is maintained by recurrent mutation, considers that a far from negligible portion of the variability is preserved by natural selection. These two theories, though they are not incompatible alternatives, have interestingly different theoretical and even philosophical implications.

According to the classical theory, a majority of the genes a species has are fixed, uniform, and homozygous in most, or even in all, individuals. The minority of the genes that are not fixed are represented by two or more variant alleles. One allele of each gene is "normal," and it confers a high fitness on its possessors; the others are mutant, abnormal, and more or less deleterious. This view is logically connected with the conception of the "optimal" genotype. The optimal genotype is the one that contains only normal genes. Any deviation from the optimum genotype will be detrimental. Nevertheless, the process of mutation unrelentingly generates detrimental mutant genes, lowers fitness, and creates a genetic load. A genetic load can, in this view, be defined as the deviation of the observed fitness from that produced by the optimum genotype. The weakness of this definition is obvious: the optimum genotype is operationally elusive; it cannot be located and its fitness cannot be measured.

The balance theory of population structure

The balance theory of population structure takes a somewhat more optimistic view. Not all genetic variability present in populations is a regrettable departure from the optimal genotype. Some of this variability is, on the contrary, an adaptive device that permits the population to secure a firmer hold on its environment. The environment is never uniform; it has a greater or lesser variety of ecological niches, opportunities for living that members of the population can exploit. A living species or population faces many diverse environments, not a single environment. Two genetic strategies are possible to cope with environmental complexity. First, there may be genotypes that react favourably in several environments. Second, there may be a variety of more or less specialized genotypes to match the variety of environments.

Nature has used both strategies in the evolutionary process. A sexually reproducing population contains a variety of hereditary endowments. Most, and perhaps all, individuals may be heterozygous for many genes; it is now known, for example, that humans are heterozygous at a minimum of 30 percent of the genetic loci responsible for the serum proteins. Any gene that is rare will be carried in heterozygotes much more frequently than in homozygotes (except in highly inbred or self-fertilizing species). Natural selection operates in sexual populations chiefly with heterozygotes. It will tend to make the heterozygotes highly fit, even if the corresponding homozygotes are low in fitness. In other words, natural selection promotes hybrid vigour or heterosis and maintains the variant genes in a state of balanced polymorphism. It takes a variety of genetic endowments to make a world worth living in.

Natural selection cannot prevent the appearance of some individuals of inferior fitness. In a population in which no two individuals carry the same genes, it is not possible to describe the genetic load as the deviation from the optimal genotype. This latter will be represented by a single individual or not at all. An operationally meaningful measure of fitness is that of the average or norm for a given population. The pronounced negative deviants from this average may be considered as constituents of the genetic load and the positive deviants as the genetic elite. Between the genetic load and the genetic elite there is the adaptive norm comprising a majority of the genotypes that are formed. The division lines between the load, the adaptive norm, and the elite can only be established by a convention. For most purposes, the load and the elite should probably refer to fairly extreme negative or positive deviations from the population mean. One may, however, foresee that some genetic processes will cause these deviations to become more or less frequent. For example, a genetic radiation damage or a greater incidence of marriages between relatives would increase the expressed genetic load compared to what it is at present.

The most impressive success scored to date, in attempts to utilize the genetic elite of an agricultural plant for increasing yields of a useful product, is represented by hybrid corn. Hybrid corn is obtained by artificially crossing inbred corn strains in order to exploit the phenomenon of hybrid vigour or heterosis.

ARTIFICIAL SELECTION

Hybrid corn is the most impressive but not the only achievement of genetics in agriculture. General accounts of animal and plant breeding can be found in the article FARMING. This section is concerned only with the genetic basis of the methods used in breeding programs. It should be kept in mind that the inbreeding-heterosis technique is most likely to succeed with species that are normally cross-fertilizing and outbred and less so in habitually inbreeding or self-pollinating forms (like wheat and barley).

In practicing artificial selection, the breeder seeks to identify and propagate those genotypes that, when placed in suitable and economically feasible environments, induce in their carriers the desired qualities. The improvement methods used in the breeding work must evidently be appropriate to the reproductive biology, the genetic population structure, and the economic value of the form to be improved. It is evidently impossible, for example, with large domestic animals such as cattle or horses to raise numbers of individuals from which selection is made comparable to that practicable with field crops. Collections of pedigrees, published in stud books for horses and in herd books for some breeds of cattle, are still regarded by commercial breeders as important in the choice of most prized animals. Progeny and sib testing are often used as aids for evaluation of the breeding worth of individual animals. A quite different situation is encountered with some useful plants in which cross-pollination happens only rarely, and the normal method of propagation is self-pollination (as in most cereals, peas, beans, and tomatoes) or parthenogenesis or asexual reproduction (sugarcane, bluegrass, bananas, and many citrus fruits). Here the problem is to identify the individual, or a small group of individuals, with a desired combination of traits and to propagate them to form pure lines (progenies obtained by self-pollination) or clones (progenies obtained asexually). Some of the most valuable varieties of wheats are descended from a single seed progenitor.

Selection, artificial as well as natural, can operate only if genetic diversity exists among the materials available. In a genetically uniform pure line or a clone, selection is without effect. Provision of genetic variability is the prime concern for a breeder. A powerful means of inducing genetic variability is hybridization and Mendelian segregation in the hybrid progenies. Two lines, neither of which is particularly good, may produce valuable genotypes in segregating hybrid progenies. This is the reason that modern breeders frequently endeavour to collect primitive varieties from far-off lands and use them as materials for hybridization followed by selection. Many, if not most, breeds and varieties, from thoroughbred horses to cultivated strawberries, are descended from hybrids of two or more ancient breeds or even of distinct species. The cultivated roses have genes from at least four wild ancestral species in their gene pool.

A technique to speed up the improvement of useful plants and animals is induction of mutations that increase the amount of genetic variability available for selection. Although a decided majority of mutants are harmful, some exceptional ones may be useful. (T.D./Ar.R.)

EUGENICS

Breeders have been quite successful in improving, for human benefit, the genetic endowments of domesticated animals, cultivated plants, and even microorganisms. The question inevitably presents itself whether humans can succeed equally well in directing the evolution of their own species toward goals regarded as good and desirable. The term eugenics has been applied to theories and practices ostensibly designed to improve the human condition from the genetic point of view. The intention of so-called negative eugenics is to improve the human species by identifying individuals and couples at risk of perpetuating "inferior" genes and to prevent such persons from reproducing. On the other hand, positive eugenics attempts to identify "good" genes and thereby improve the species by encouraging the reproduction of those persons who are thought to carry them. All too often, however, the identification of desirable or undesirable hereditary human traits is a subjective and controversial matter, in which the opinions of one, often self-serving, group become the basis of reproductive decisions imposed upon another group. Furthermore, the concept of eugenics tends to ignore the sizable role that environment plays in the establishment of human characteristics.

The idea of applying knowledge of heredity to the improvement of the human race goes back to earliest times. References to eugenic ideals appear in the Old Testament, and Plato's *Republic* idealizes a society in which there is constant selection for the improvement of the human stock. A British economist, Thomas Robert Malthus, noted the struggle for existence; Darwin saw in it the means of evolution. It remained for Francis Galton, a scientist of many talents and a cousin of Darwin, to see that the theory of evolution implied that humans might in part direct their own evolutionary future. Galton's first important work, *Hereditary Genius* (1869), contained the results of his studies of the families of eminent men as evidence for his belief that "it would be quite practical to produce a highly gifted race of men by judicious marriages during several consecutive generations." In 1883 he published *Inquiries into Human Faculty,* in which he coined the term eugenics. In *Natural Inheritance* (1889), Galton pioneered the development and application of advanced statistical methods to the study of man.

While Galton's thinking was much ahead of his time, he retained some of the prejudices of an English gentleman of his day in regard to social class and race. In his studies of the families of eminent men, Galton underestimated the effects of a favourable environment. But his faults in this respect were far less than those of many of his followers. He developed many of the statistical methods for the study of populations and was the first to recognize the value of the study of twins for research in heredity. He repudiated the Lamarckian view, then widely held, that acquired characteristics were inherited. Although Galton attached religious significance to the eugenic movement, he did not think of it in revolutionary terms but rather as "the new duty which is supposed to be exercised concurrently with and not in opposition to, the old ones upon which the social fabric depends." For all his ability, Galton was not able to gain wide acceptance for eugenics, largely because much of the scientific and technical foundation was lacking.

Galton had endowed a research fellowship in eugenics in 1904 and, in his will, provided funds for a chair of eugenics at University College, London University. The fellowship and later the chair at University College were both occupied by Karl Pearson, a brilliant mathematician who helped to create the science of biometry, the statistical aspects of biology. Pearson was a controversial figure who believed that environment had little to do with the development of mental or emotional qualities. He felt that the high birth rate of the poor was a threat to civilization and that the "higher" races must supplant the "lower." His views gave countenance to those who believed in racial and class superiorities. Thus, Pearson shares the blame for the discredit later brought on eugenics in the United States and for making possible the dreadful misuse of the word eugenics in Adolf Hitler's propaganda. The English Eugenics Society, founded by Galton in 1907 as the Eugenics Education Society, opposed Pearson's views but was itself slow in throwing off the prevalent biases of that time.

Eugenics discredited by early racist views

In the United States the American Eugenics Society was founded in 1926 by men who believed that the white race was superior to other races and, further, that the "Nordic" white was superior to other whites. They thought of races as "pure" groups, quite separate from each other. They did not know that all races are mixtures of many types, the distribution of genes among the races varying in proportions rather than in kind, as evidenced by the distribution of the various blood groups among all of the races. They did not realize that environments are so uncontrollable that scientists cannot reasonably put forward views on the genetic differences in the performance of different races. Furthermore, they believed that the upper classes had superior hereditary qualities that justified their being the ruling class.

The science of that day supported extreme views on "feeblemindedness" and "criminal types." Intelligence tests, introduced in the early 1900s by the French psychologist Alfred Binet, were thought, contrary to Binet's views, to be measures of innate, genetic intelligence. People whose test performance (*e.g.,* intelligence quotient, or IQ) gave them a mental age of 12 or less were classified as feebleminded or morons. Criminality was generally considered a concomitant of feeblemindedness. Studies on degenerate family stocks were taken to prove that hundreds of persons in each of these families were feebleminded or criminal types because of the inheritance they received from a single ancestor five or six generations back. Claims were made that immigrants from southern and eastern Europe, besides being socially inferior, included criminal and defective stocks. There was much in the intellectual atmosphere of the United States that fostered an extreme hereditarian racist view.

Even before 1905 eugenists had been active in the effort to restrict immigration. Their arguments, along with those of others who advocated restrictions for different reasons, culminated with the passage of the Immigration Act of 1924, which limited quota immigrants to about 150,000 annually, with no more than 2 percent of each nationality according to the number of persons of their national origin in the United States as of 1890, and providing that, beginning July 1, 1929, the quota of any country should have the same ratio to 150,000 as the number of persons of that national origin living in the United States had to the total U.S. population in the 1920 census. In later years it became clear that the material the eugenists had presented to congressional hearings had little scientific foundation.

Eugenists at this time laid great stress on the importance of sterilizing defective persons. By 1931 sterilization laws had been enacted by 27 states in the United States, and by 1935 sterilization laws had been passed in Denmark, Switzerland, Germany, Norway, and Sweden. Most of these laws provided for the voluntary or compulsory sterilization of certain classes of people thought to be insane, mentally retarded, or epileptic; some applied equally to habitual criminals and sexual deviants. In most cases the purpose was clearly eugenic, though some laws tacitly permitted sterilization for social rather than genetic reasons. In the United States most states did not generally enforce these rather extreme measures, and the number of sterilizations was seldom more than 100 per year. Exceptions were California, where sterilizations averaged more than 350 cases per year, with a total of 9,931 by 1935, and some of the southern states, with fairly high sterilization rates relative to their populations.

Unquestionably the greatest abuse of the concept of eugenics took place in Hitler's Germany, when as a rationale for producing a "master race" the Nazis murdered millions of people considered to have inferior genes.

As a result of these misuses, modern society has been disinclined to admit any validity to eugenic concepts. However, the practice of modern genetic counseling (see below) is in a special sense a eugenic activity, in that it attempts to prevent the conception or birth of individuals with most serious forms of maldevelopment that would be tragic burdens to themselves and to their families. This form of negative eugenics identifies individuals and couples at risk of perpetuating genes that lead to heritable diseases and disorders that contribute to a significant amount of human tragedy. Of importance is the fact that information on these risks is given to couples so that they can make informed and personal decisions about reproduction without societal pressure. (F.H.O./Ar.R.)

HUMAN GENETICS

The study of human heredity occupies a central position in genetics. Much of this interest stems from a basic desire to know who humans are and why they are as they are. At a more practical level, an understanding of human heredity is of critical importance in the prediction, diagnosis, and treatment of diseases that have a genetic component. The quest to determine the genetic basis of human health has given rise to the field of medical genetics. In general, medicine has given focus and purpose to human genetics, so that the terms medical genetics and human genetics are often considered synonymous.

General aspects

THE HUMAN CHROMOSOMES

A new era in cytogenetics, the field of investigation concerned with studies of the chromosomes, began in 1956 with the discovery by Jo Hin Tjio and Albert Levan that human somatic cells contain 23 pairs of chromosomes. Since that time the field has advanced with amazing rapidity and has demonstrated that human chromosome aberrations rank as major causes of fetal death and of tragic human diseases, many of which are accompanied by mental retardation. Since the chromosomes can be delineated only during mitosis, it is necessary to examine material in which there are many dividing cells. This can usually be accomplished by culturing cells from the blood or skin, since only the bone marrow cells (not readily sampled except during serious bone marrow disease such as leukemia) have sufficient mitoses in the absence of artificial culture. After growth, the cells are fixed on slides and then stained with a variety of DNA-specific stains that permit the delineation and identification of the chromosomes. The Denver system of chromosome classification, established in 1959, identified the chromosomes by their length and the position of the centromeres. Since then the method has been improved by the use of special staining techniques that impart unique light and dark bands to each chromosome. These bands permit the identification of chromosomal regions that are duplicated, missing, or transposed to other chromosomes.

Figure 15 shows the karyotypes (*i.e.,* the physical appearance of the chromosomes) of a male and female. In these micrographs the 46 human chromosomes (the diploid number) are arranged in homologous pairs, each consisting of one maternally derived and one paternally derived member. The chromosomes are all numbered except for the X and the Y chromosomes, which are the sex chromosomes. In humans, as in all mammals, the normal female has two X chromosomes and the normal male has one X chromosome and one Y chromosome. The female is thus the homogametic sex, as all her gametes normally have one X chromosome. The male is heterogametic, as he produces two types of gametes—one type containing an X chromosome and the other containing the Y chromosome. There is good evidence that the Y chromosome in humans, unlike that in *Drosophila,* is necessary (but not sufficient) for maleness.

FERTILIZATION, SEX DETERMINATION, AND DIFFERENTIATION

A human individual arises through the union of two cells, an egg from the mother and a sperm from the father. Human egg cells are barely visible to the naked eye. They are shed, usually one at a time, from the ovary into the oviducts (fallopian tubes), through which they pass into the uterus. Fertilization, the penetrance of an egg by a sperm, occurs in the oviducts. This is the main event of sexual reproduction and determines the genetic constitution of the new individual.

Human sex determination is a genetic process that depends basically on the presence of the Y chromosome in the fertilized egg. This chromosome stimulates a change in the undifferentiated gonad into that of the male (a testicle). The gonadal action of the Y chromosome is mediated by a gene located near the centromere; this gene codes for the production of a cell surface molecule called the H-Y antigen. Further development of the anatomic structures, both internal and external, that are associated with maleness is controlled by hormones produced by the testicle. The sex of an individual can be thought of in three different contexts: chromosomal sex, gonadal sex, and anatomic sex. Discrepancies among these, especially the latter two, result in the development of individuals with ambiguous sex, often called hermaphrodites. Homosexuality is of uncertain cause and is unrelated to the above sex-determining factors. It is of interest that in the absence of a male gonad (testicle) the internal and external sex anatomy is always female, even in the absence of a female ovary. A female without ovaries will, of course, be infertile and will not experience any of the female developmental changes normally associated with puberty. Such a female will often have Turner's syndrome (see below).

If X-containing and Y-containing sperm are produced in equal numbers, then according to simple chance one would expect the sex ratio at conception (fertilization) to be half boys and half girls, or 1 : 1. Direct observation of sex ratios among newly fertilized human eggs is not yet feasible, and sex-ratio data are usually collected at the time of birth. In almost all human populations of newborns there is a slight excess of males; about 106 boys are born for each 100 girls. Throughout life, however, there is a slightly greater mortality of males; this slowly alters the sex ratio until, beyond the age of about 50 years, there is an excess of females. Studies indicate that male embryos suffer a relatively greater degree of prenatal mortality, so that the sex ratio at conception might be expected to favour males even more than the 106 : 100 ratio observed at birth would suggest. Firm explanations for the apparent excess of male conceptions have not been established; it is possible that Y-containing sperm survive better within the female reproductive tract, or that they may be a little more successful in reaching the egg in order to fertilize it. In any case, the sex differences are small, the statistical expectation for a boy (or girl) at any single birth still being close to one out of two.

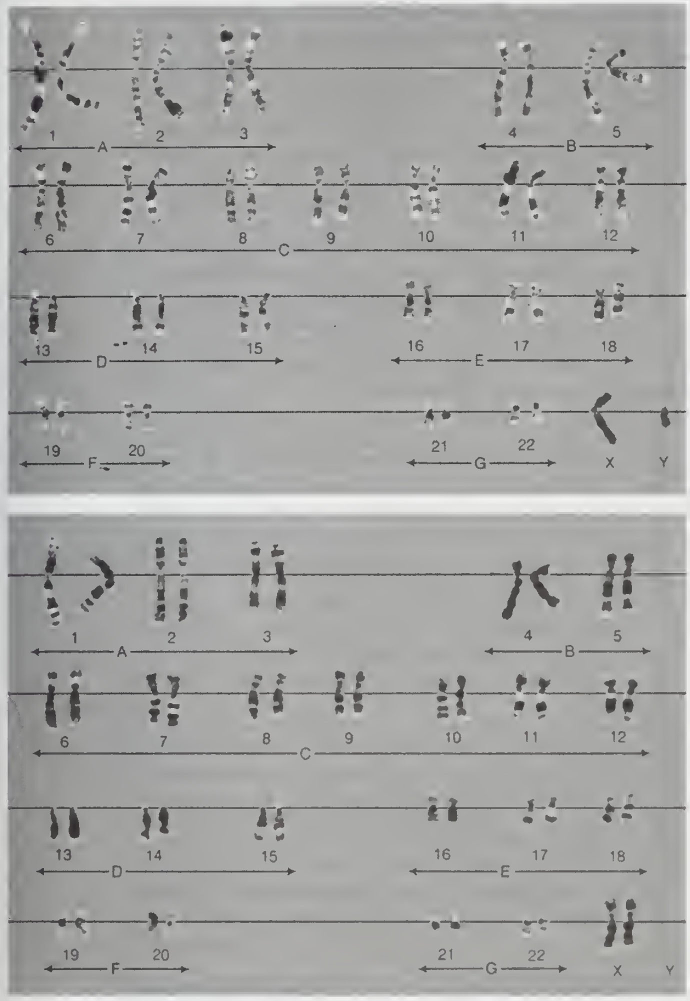


Figure 15: *Human karyotypes of the (top) male and female.* The chromosomes have been stained with Giemsa stain. Individual chromosomes are characterized by their size, position of centromere, and unique distribution of light and dark bands.

From the Cytogenetics Laboratory of Dr. Arthur Robinson, National Jewish Hospital and Research Center/National Asthma Center Denver, Colorado

During gestation—the period of nine months between fertilization and the birth of the infant—a remarkable series of developmental changes occur. Through the process of mitosis, the total number of cells changes from one (the fertilized egg) to about 2×10^{11}. In addition, these cells differentiate into hundreds of different types with specific functions (liver cells, nerve cells, muscle cells, etc.). A multitude of regulatory processes, both genetically and environmentally controlled, accomplish this differentiation. Elucidation of the exquisite timing of these processes remains one of the great challenges of human biology.

IMMUNOGENETICS

Immunity is the ability of an individual to recognize the "self" molecules that make up one's own body and to distinguish them from such "non-self" molecules as those found in infectious microorganisms and toxins. This process has a prominent genetic component. Knowledge of the genetic and molecular basis of the mammalian immune system has increased in parallel with the explosive advances made in somatic cell and molecular genetics.

Lymphocytes

There are two major components of the immune system, both originating from the same precursor "stem" cells. The bursa component provides B lymphocytes, a class of white blood cells that, when appropriately stimulated, differentiate into plasma cells. These latter cells produce circulating soluble proteins called antibodies or immunoglobulins. Antibodies are produced in response to substances called antigens, most of which are foreign proteins or polysaccharides. An antibody molecule can recognize a specific antigen, combine with it, and initiate its destruction. This so-called humoral immunity is accomplished through a complicated series of interactions with other molecules and cells; some of these interactions are mediated by another group of lymphocytes, the T lymphocytes, which are derived from the thymus gland. Once a B lymphocyte has been exposed to a specific antigen, it "remembers" the contact so that future exposure will cause an accelerated and magnified immune reaction. This is a manifestation of what has been called immunological memory.

The thymus component of the immune system centres on the thymus-derived T lymphocytes. In addition to regulating the B cells in producing humoral immunity, the T cells also directly attack cells that display foreign antigens. This process, called cellular immunity, is of great importance in protecting the body against a variety of viruses as well as cancer cells. Cellular immunity is also the chief cause of the rejection of organ transplants. The T lymphocytes provide a complex network consisting of a series of helper cells (which are antigen specific), amplifier cells, suppressor cells, and cytotoxic (killer) cells, all of which are important in immune regulation.

The genetics of antibody formation. One of the central problems in understanding the genetics of the immune system has been in explaining the genetic regulation of antibody production. Immunobiologists have demonstrated that the system can produce well over 1,000,000 specific antibodies, each corresponding to a particular antigen. It would be difficult to envisage that each antibody is encoded by a separate gene—such an arrangement would require a disproportionate share of the entire human genome. Recombinant DNA analysis has illuminated the mechanisms by which a limited number of immunoglobulin genes can encode this vast number of antibodies.

Each antibody molecule consists of several different polypeptide chains—the light chains (L) and the longer heavy chains (H). The latter determine to which of five different classes (IgM, IgG, IgA, IgD, or IgE) an immunoglobulin belongs. Both the L and H chains are unique among proteins in that they contain constant and variable parts. The constant parts have relatively identical amino acid sequences in any given antibody. The variable parts, on the other hand, have different amino acid sequences in each antibody molecule. It is the variable parts, then, that determine the specificity of the antibody.

Recombinant DNA studies of immunoglobulin genes in mice have revealed that the light-chain genes are encoded in four separate parts in germ-line DNA: a leader segment (L), a variable segment (V), a joining segment (J), and a constant segment (C). These segments are widely separated in the DNA of an embryonic cell, but in a mature B lymphocyte they are found in relative proximity (albeit separated by introns). The mouse has more than 200 light-chain variable region genes, only one of which will be incorporated into the proximal sequence that codes for the antibody production in a given B lymphocyte. Antibody diversity is greatly enhanced by this system, as the V and J segments rearrange and assort randomly in each B-lymphocyte precursor cell. The mechanisms by which this DNA rearrangement takes place are not clear, but transposons are undoubtedly involved. Similar combinatorial processes take place in the genes that code for the heavy chains; furthermore, both the light-chain and heavy-chain genes can undergo somatic mutations to create new antibody-coding sequences. The net effect of these combinatorial and mutational processes enables the coding of millions of specific antibody molecules from a limited number of genes. It should be stressed, however, that each B lymphocyte can produce only one antibody. It is the B lymphocyte population as a whole that produces the tremendous variety of antibodies in humans and other mammals.

Plasma cell tumours (myelomas) have made it possible to study individual antibodies since these tumours, which are descendants of a single plasma cell, produce one antibody in abundance. Another method of obtaining large amounts of a specific antibody is by fusing a B lymphocyte with a rapidly growing cancer cell. The resultant hybrid cell, known as a hybridoma, multiplies rapidly in culture. Since the antibodies obtained from hybridomas are produced by clones derived from a single lymphocyte, they are called monoclonal antibodies.

The genetics of cellular immunity. As has been stated, cellular immunity is mediated by T lymphocytes that can recognize infected body cells, cancer cells, and the cells of a foreign transplant. The control of cellular immune reactions is provided by a linked group of genes, known as the major histocompatibility complex (MHC). These genes code for the major histocompatibility antigens, which are found on the surface of almost all nucleated somatic cells. The major histocompatibility antigens were first discovered on the leukocytes (white blood cells) and are, therefore, usually referred to as the HLA (human leukocyte group A) antigens.

The advent of the transplantation of human organs in the 1950s made the question of tissue compatibility between donor and recipient of vital importance, and it was in this context that the HLA antigens and the MHC were elucidated. Investigators found that the MHC resides on the short arm of chromosome 6, on four closely associated sites designated HLA-A, HLA-B, HLA-C, and HLA-D. Each locus is highly polymorphic—*i.e.,* each is represented by a great many alleles within the human gene pool. These alleles, like those of the ABO blood group system, are expressed in codominant fashion. Because of the large number of alleles at each HLA locus, there is an extremely low probability of any two individuals (other than siblings) having identical HLA genotypes. (Since a person inherits one chromosome 6 from each parent, siblings have a 25 percent probability of having received the same paternal and maternal chromosomes 6 and thus of being HLA matched.)

Although HLA antigens are largely responsible for the rejection of organ transplants, it is obvious that the MHC did not evolve to prevent the transfer of organs from one person to another. Indeed, information obtained from the histocompatibility complex in the mouse (which is very similar in its genetic organization to that of the human) suggests that a primary function of the HLA antigens is to regulate the number of specific cytotoxic T killer cells, which have the ability to destroy virus-infected cells and cancer cells. (Ar.R.)

THE GENETICS OF HUMAN BLOOD

More is known about the genetics of the blood than about any other human tissue. One reason for this is that blood samples can be easily secured and subjected to biochemical analysis without harm or major discomfort to the person being tested. Perhaps a more cogent reason is that many chemical properties of human blood display relatively simple patterns of inheritance.

Blood types. Certain chemical substances within the red blood cells (such as the ABO and MN substances noted above) may serve as antigens. When cells that contain specific antigens are introduced into the body of an experimental animal such as a rabbit, the animal responds by producing antibodies in its own blood.

In addition to the ABO and MN systems, geneticists have identified about 14 blood-type gene systems associated with other chromosomal locations. The best known of these is the Rh system. The Rh antigens are of particular importance in human medicine. Curiously, however, their existence was discovered in monkeys. When blood from the rhesus monkey (hence the designation Rh) is injected into rabbits, the rabbits produce so-called Rh antibodies that will agglutinate not only the red blood cells of the monkey but the cells of a large proportion of human beings as well. Some people (Rh-negative individuals), however, lack the Rh antigen; the proportion of such persons varies from one human population to another. Akin to data concerning the ABO system, the evidence for Rh genes indicates that only a single chromosome locus (called *r*) is involved and is located on chromosome 1. At least 35 Rh alleles are known for the *r* location; basically the Rh-negative condition is recessive.

Rh incompatibility

A medical problem may arise when a woman who is Rh-negative carries a fetus that is Rh-positive. The first such child may have no difficulty, but later similar pregnancies may produce severely anemic newborn infants. Exposure to the red blood cells of the first Rh-positive fetus appears to immunize the Rh-negative mother, that is, she develops antibodies that may produce permanent (sometimes fatal) brain damage in any subsequent Rh-positive fetus. Damage arises from the scarcity of oxygen reaching the fetal brain because of the severe destruction of red blood cells. Measures are available for avoiding the severe effects of Rh incompatibility by transfusions to the fetus within the uterus; however, genetic counselling before conception is helpful so that the mother can receive Rh immunoglobulin immediately after her first and any subsequent pregnancies involving an Rh-positive fetus. This immunoglobulin effectively destroys the fetal red blood cells before the mother's immune system is stimulated. The mother thus avoids becoming actively immunized against the Rh antigen and will not produce antibodies that could attack the red blood cells of a future Rh-positive fetus.

Serum proteins. Human serum, the fluid portion of the blood that remains after clotting, contains various proteins that have been shown to be under genetic control. Study of genetic influences has flourished since the development of precise methods for separating and identifying serum proteins. These move at different rates under the impetus of an electrical field (electrophoresis), as do proteins from many other sources (*e.g.,* muscle or nerve). Since the composition of a protein is specified by the structure of its corresponding gene, biochemical studies based on electrophoresis permit direct study of tissue substances that are only a metabolic step or two away from the genes themselves.

Electrophoretic studies have revealed that at least one-third of the human serum proteins occur in variant forms. Many of the serum proteins are polymorphic, occurring as two or more variants with a frequency of not less than 1 percent each in a population. Patterns of polymorphic serum protein variants have been used to determine whether twins are identical (as in assessing compatibility for organ transplants) or whether two individuals are related (as in resolving paternity suits). Whether or not the different forms have a selective advantage is not generally known.

Much attention in the genetics of substances in the blood has been centred on serum proteins called haptoglobins, transferrins (which transport iron), and gamma globulins (a number of which are known to immunize against infectious diseases). Haptoglobins appear to relate to two common alleles at a single chromosome locus; the mode of inheritance of the other two seems more complicated, about 18 kinds of transferrins having been described. Like blood-cell antigen genes, serum-protein genes are distributed worldwide in the human population in a way that permits their use in tracing the origin and migration of different groups of people.

Hemoglobin. Hundreds of variants of hemoglobin have been identified by electrophoresis, but relatively few are frequent enough to be called polymorphisms. Of the polymorphisms, the alleles for sickle-cell and thalassemia hemoglobins produce serious disease in homozygotes, whereas others (hemoglobins C, D, and E) do not. As was discussed earlier (see above *Varieties of natural selection*), the sickle-cell polymorphism confers a selective advantage on the heterozygote living in a malarial environment; the thalassemia polymorphism provides a similar advantage.

INFLUENCE OF THE ENVIRONMENT

As stated earlier in this article, gene expression occurs only after modification by the environment. A good example is the recessively inherited disease called galactosemia, in which the enzyme necessary for the metabolism of galactose—a component of milk sugar—is defective. The sole source of galactose in the infant's diet is milk, which in this instance is toxic. The treatment of this most serious disease in the neonate is to remove all natural forms of milk from the diet (environmental manipulation) and to substitute a synthetic milk lacking galactose. The infant will then develop normally but will never be able to tolerate foods containing lactose. If milk were not a major part of the infant's diet, however, the mutant gene would never be able to express itself, and galactosemia would be unknown.

The interaction of heredity and environment

Another way of saying this is that no trait can exist or become actual without an environmental contribution. Thus, the old question of which is more important, heredity or environment, is without meaning. Both nature (heredity) and nurture (environment) are always important for every human attribute.

But this is not to say that the separate contributions of heredity and environment are equivalent for each characteristic. Dark pigmentation of the iris of the eye, for example, is under hereditary control in that one or more genes specify the synthesis and deposition in the iris of the pigment (melanin). This is one character that is relatively independent of such environmental factors as diet or climate; thus, individual differences in eye colour tend to be largely attributable to hereditary factors rather than to ordinary environmental change.

On the other hand, it is unwarranted to assume that other traits (such as height, weight, or intelligence) are as little affected by environment as is eye colour. It is very easy to gather information that tall parents tend, on the average, to have tall children (and that short parents tend to produce short children), properly indicating a hereditary contribution to height. Nevertheless, it is equally manifest that growth can be stunted in the environmental absence of adequate nutrition. The dilemma arises that only the combined, final result of this nature–nurture interaction can be directly observed. There is no accurate way (in the case of a single individual) to gauge the separate contributions of heredity and environment to such a characteristic as height. An inferential way out of this dilemma is provided by studies of twins.

Fraternal twins. Usually a fertile human female produces a single egg about once a month. Should fertilization occur (a zygote is formed), growth of the individual child normally proceeds after the fertilized egg has become implanted in the wall of the uterus (womb). In the unusual circumstance that two unfertilized eggs are simultaneously released by the ovaries, each egg may be fertilized by a different sperm cell at about the same time, become implanted, and grow, to result in the birth of twins.

Twins formed from separate eggs and different sperm cells can be of the same or of either sex. No matter what their sex, they are designated as fraternal twins. This terminology is used to emphasize that fraternal twins are genetically no more alike than are siblings (brothers or sisters) born years apart. Basically they differ from ordinary siblings only in having grown side by side in the womb and in having been born at approximately the same time.

Split embryos

Identical twins. In a major nonfraternal type of twinning, only one egg is fertilized; but during the cleavage of this single zygote into two cells, the resulting pair somehow become separated. Each of the two cells may implant in the uterus separately and grow into a complete, whole individual. In laboratory studies with the zygotes of many animal species, it has been found that in the two-cell stage (and later) a portion of the embryo, if separated under the microscope by the experimenter, may develop into a perfect, whole individual. Such splitting occurs spontaneously at the four-cell stage in some organisms (*e.g.*, the armadillo) and has been accomplished experimentally with the embryos of salamanders, among others.

The net result of splitting at an early embryonic stage may be to produce so-called identical twins. Since such twins derive from the same fertilized egg, the hereditary material from which they originate is absolutely identical in every way, down to the last gene locus. While developmental and genetic differences between one "identical" twin and another still may arise through a number of processes (*e.g.*, mutation), these twins are always found to be of the same sex. They are often breathtakingly similar in appearance, frequently down to very fine anatomic and biochemical details (although their fingerprints are differentiable).

Diagnosis of twin types. Since the initial event in the mother's body (either splitting of a single egg or two separate fertilizations) is not observed directly, inferential means are employed for diagnosing a set of twins as fraternal or identical. The birth of fraternal twins is frequently characterized by the passage of two separate afterbirths. In many instances, identical twins are followed by only a single afterbirth, but exceptions to this phenomenon are so common that this is not a reliable method of diagnosis.

The most trustworthy method for inferring twin type is based on the determination of genetic similarity. By selecting those traits that display the least variation attributable to environmental influences (such as eye colour and blood types), it is feasible, if enough separate chromosome loci are considered, to make the diagnosis of twin type with high confidence.

Inferences from twin studies. *Metric (quantitative) traits.* By measuring the heights of a large number of ordinary siblings (brothers and sisters) and of twin pairs, it may be shown that the average difference between identical twins is less than half the difference for all other siblings (see Figure 16). Any average differences between groups of identical twins are attributable with considerable confidence to the environment. Thus, since the sample of identical twins (given in Figure 16) who were reared apart (in different homes) differed little in height from identicals who were raised together, it appears that environmental–genetic influences on that trait tended to be similar for both groups.

Twins reared together and apart

Yet, the data for like-sexed fraternal twins reveal a much greater average difference in height (about the same as that found for ordinary siblings reared in the same home at different ages). Apparently the fraternal twins were more dissimilar than identicals (even though reared together) because the fraternals differed more among themselves in genotype. This emphasizes the great genetic similarity among identicals. Such studies can be particularly enlightening when the effects of individual genes are obscured or distorted by the influence of environmental factors on quantitative (measurable) traits (*e.g.*, height, weight, and intelligence).

Any trait that can be objectively measured among identical and fraternal twins can be scrutinized for the particular combination of hereditary and environmental influences that impinge upon it. The effect of environment on identical twins reared apart (Figure 16) is suggested by their relatively great average difference in body weight as compared with identical twins reared together. Weight appears to be more strongly modified by environmental variables than is height.

Study of comparable characteristics among farm animals and plants suggests that such quantitative human traits as height and weight are affected by allelic differences at a number of chromosome locations; they are not simply affected by genes at a single locus. Investigation of these gene systems with multiple locations (polygenic systems) is carried out largely through selective-breeding experiments among large groups of plants and lower animals. Human beings select their mates in a much freer fashion, of course, and polygenic studies among people are thus severely limited.

Intelligence

Intelligence is a very complex human trait, the genetics of which has been a subject of controversy for some time. Much of the controversy arises from the fact that intelligence is so difficult to define. Information has been based almost entirely on scores on standardized IQ tests constructed by psychologists; in general such tests do not take into account cultural, environmental, and educational differences. As a result, the working definition of intelligence has been "the general factor common to a large number of diverse cognitive (IQ) tests." Even roughly measured as IQ, intelligence shows (Figure 16) a strong contribution from the environment. Fraternal twins, however, show relatively great dissimilarity in IQ, suggesting an important contribution from heredity as well. In fact, it has been estimated that on the average between 60 and 80 percent of the variance in IQ test scores could be genetic. It is important to note that intelligence is polygenically inherited and that it has the highest degree of assortative mating of any trait; in other words, people tend to mate with people having similar IQ's. Moreover, twin studies involving psychological traits should be viewed with caution; for example, since identical twins tend to be singled out for special attention, their environment should not be considered equivalent even to that of other children raised in their own family.

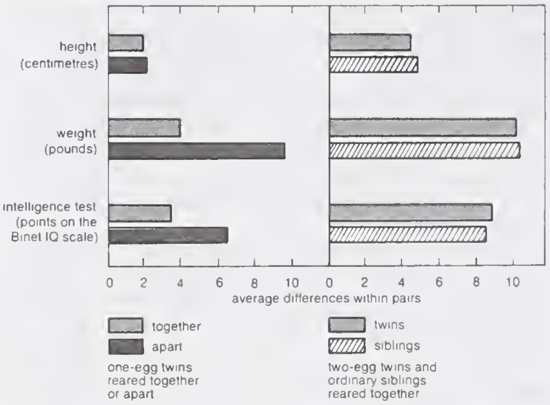


Figure 16: Measured differences between identical (one-egg) twins, between like-sexed fraternal (two-egg) twins, and between like-sexed ordinary siblings.

Adapted from H.L. Carson, *Heredity and Human Life* (1963), Columbia University Press

Other traits. For traits of a more qualitative (all-or-none) nature, the twin method can also be used in efforts to assess the degree of hereditary contribution. Such investigations are based on an examination of cases in which at least one member of the twin pair shows the trait. It was found in one study, for example, that in about 80 percent of all identical twin pairs in which one twin shows symptoms of the psychiatric disorder called schizophrenia, the other member of the pair also shows the symptoms (that is, the two are concordant for the schizophrenic trait). In the remaining 20 percent, the twins are discordant (that is, one lacks the trait). Since identical twins often have similar environments, this information by itself does not distinguish between the effects of heredity and environment. When pairs of like-sexed fraternal twins reared together are studied, however, the degree of concordance for schizophrenia is very much lower—only about 15 percent (Figure 17).

Schizophrenia thus clearly develops much more easily in some genotypes than among others; this indicates a strong hereditary predisposition to the development of the trait. Schizophrenia also serves as a good example of the influence of environmental factors since concordance for the condition does not appear in 100 percent of identical twins.

Studies of concordance and discordance between identical and fraternal twins have been carried out for many other human characteristics, a few of which are also summarized in Figure 17. It has, for example, been known for many years that tuberculosis is a bacterial infection of environmental origin. Yet identical twins raised in the same home show concordance for the disease far more often than do fraternal twins. This finding seems to be explained by the high degree of genetic similarity between the identical twins. While the tuberculosis germ is not inherited, heredity does seem to make one more (or less) susceptible to this particular infection. Thus, the genes of one individual may provide the chemical basis for susceptibility to a disease, while the genes of another may fail to do so.

Adapted from H.L. Carson, *Heredity and Human Life* (1963); Columbia University Press

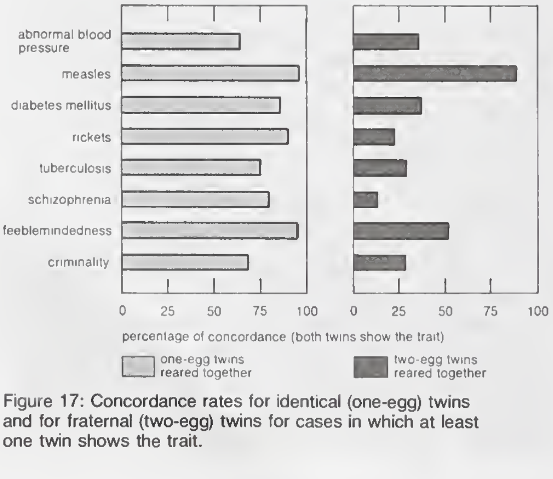


Figure 17: Concordance rates for identical (one-egg) twins and for fraternal (two-egg) twins for cases in which at least one twin shows the trait.

Indeed, there seem to be genetic differences between disease germs themselves that result in differences in their virulence. Thus, whether a genetically susceptible person actually develops a disease also depends in part on the heredity of the particular strain of bacteria or virus with which he or she must cope. Consequently, unless environmental factors such as these are adequately evaluated, the conclusions drawn from susceptibility studies can be unfortunately misleading.

The above discussion should help to make clear the limits of genetic determinism. The expression of the genotype can always be modified by the environment. It can be argued that all human illnesses have a genetic component and that the basis of all medical therapy is environmental modification. Specifically, this is the hope for the management of genetic diseases. The more that can be learned about the basic molecular and cellular dysfunctions associated with such diseases, the more amenable they will be to environmental manipulation. (H.L.C./Ar.R.)

Genetics and human disease

With the increasing ability to control infectious and nutritional diseases in developed countries, there has come the realization that genetic diseases are a major cause of disability, death, and human tragedy. Rare, indeed, is the family that is entirely free of any known genetic disorder. Many thousands of different genetic disorders with defined clinical symptoms have been identified. Of the 3 to 6 percent of newborns with a recognized birth defect, at least half involve a predominantly genetic contribution. Furthermore, genetic defects are the major known cause of pregnancy loss in developed nations, and almost half of all spontaneous abortions (miscarriages) involve a chromosomally abnormal fetus. About 30 percent of all postnatal infant mortality in developed countries is due to genetic disease; 30 percent of pediatric and 10 percent of adult hospital admissions can be traced to a predominantly genetic cause. Finally, medical investigators estimate that genetic defects—however minor—are present in at least 10 percent of all adults. Thus, these are not rare events.

A congenital defect is any biochemical, functional, or structural abnormality that originates prior to or shortly after birth. It must be emphasized that birth defects do not all have the same basis, and it is even possible for apparently identical defects in different individuals to reflect different underlying causes. Though the genetic and biochemical bases for most recognized defects are still uncertain, it is evident that many of these disorders result from a combination of genetic and environmental factors.

This article surveys the main categories of genetic disease, focusing on the types of genetic mutations that give rise to them, the risks associated with exposure to certain environmental agents, and the course of managing genetic disease through counseling, diagnosis, and treatment.

CLASSES OF GENETIC DISEASE

Most human genetic defects can be categorized as resulting from either chromosomal, single-gene Mendelian, single-gene non-Mendelian, or multifactorial causes. Each of these categories is discussed briefly below.

Diseases caused by chromosomal aberrations. About one out of 150 live newborns has a detectable chromosomal abnormality. Yet even this high incidence represents only a small fraction of chromosome mutations since the vast majority are lethal and result in prenatal death or stillbirth. Indeed, 50 percent of all first-trimester miscarriages and 20 percent of all second-trimester miscarriages are estimated to involve a chromosomally abnormal fetus.

Chromosome disorders can be grouped into three principal categories: (1) those that involve numerical abnormalities of the autosomes, (2) those that involve structural abnormalities of the autosomes, and (3) those that involve the sex chromosomes. Autosomes are the 22 sets of chromosomes found in all normal human cells. They are referred to numerically (*e.g.,* chromosome 1, chromosome 2) according to a traditional sort order based on size, shape, and other properties. Sex chromosomes make up the 23rd pair of chromosomes in all normal human cells and come in two forms, termed X and Y. In humans and many other animals, it is the constitution of sex chromosomes that determines the sex of the individual, such that XX results in a female, and XY results in a male.

Down syndrome

Numerical abnormalities. Numerical abnormalities, involving either the autosomes or sex chromosomes, are believed generally to result from meiotic nondisjunction—that is, the unequal division of chromosomes between daughter cells—that can occur during either maternal or paternal gamete formation. Meiotic nondisjunction leads to eggs or sperm with additional or missing chromosomes. Although the biochemical basis of numerical chromosome abnormalities remains unknown, maternal age clearly has an effect, such that older women are at significantly increased risk to conceive and give birth to a chromosomally abnormal child. The risk increases with age in an almost exponential manner, especially after age 35, so that a pregnant woman age 45 or older has between a 1 in 20 and 1 in 50 chance that her child will have trisomy 21 (Down syndrome), while the risk is only 1 in 400 for a 35-year-old woman and less than 1 in 1,000 for a woman under the age of 30. There is no clear effect of paternal age on numerical chromosome abnormalities.

Although Down syndrome is probably the best known and most commonly observed of the autosomal trisomies, being found in about 1 out of 800 live births, both trisomy 13 and trisomy 18 are also seen in the population, albeit at greatly reduced rates. The vast majority of conceptions involving trisomy for any of these three autosomes are nonetheless lost to miscarriage, as are all conceptions involving pure trisomy for any of the other autosomes. Similarly, monosomy for any of the autosomes is lethal in utero and therefore is not seen in the population. Because numerical chromosomal abnormalities generally result from independent meiotic events, parents who have one pregnancy with a numerical chromosomal abnormality are generally not at markedly increased risk above the general population to repeat the experience. Nonetheless, a small increased risk is generally cited for these couples to account for unusual situations, such as chromosomal translocations or gonadal mosaicism, described below.

Structural abnormalities. Structural abnormalities of the autosomes are even more common in the population than are numerical abnormalities and include translocations of large pieces of chromosomes, as well as smaller deletions, insertions, or rearrangements. Indeed, about 5 percent of all cases of Down syndrome result not from classic trisomy 21 but from the presence of excess chromosome 21 material attached to the end of another chromosome as the result of a translocation event. If balanced, structural chromosomal abnormalities may be compatible with a normal phenotype, although unbalanced chromosome structural abnormalities can be every bit as devastating as numerical abnormalities. Furthermore, because many structural defects are inherited from a parent who is a balanced carrier, couples who have one pregnancy with a structural chromosomal abnormality generally are at significantly increased risk above the general population to repeat the experience. Clearly, the likelihood of a recurrence would depend on whether a balanced form of the structural defect occurs in one of the parents.

Even a small deletion or addition of autosomal material—too small to be seen by normal karyotyping methods—can produce serious malformations and mental retardation. One example is *cri du chat* (French: "cry of the cat") syndrome, which is associated with the loss of a small segment of the short arm of chromosome 5. Newborns with this disorder have a "mewing" cry like that of a cat. Mental retardation is usually severe.

Turner's syndrome

Abnormalities of the sex chromosomes. About 1 in 400 male and 1 in 650 female live births demonstrate some form of sex chromosome abnormality, although the symptoms of these conditions are generally much less severe than are those associated with autosomal abnormalities. Turner's syndrome is a condition of females who, in the classic form, carry only a single X chromosome (45,X). Turner's syndrome is characterized by a collection of symptoms, including short stature, webbed neck, and incomplete or absent development of secondary sex characteristics, leading to infertility. Although Turner's syndrome is seen in about 1 in 2,500 to 1 in 5,000 female live births, the 45,X karyotype accounts for 10 to 20 percent of the chromosomal abnormalities seen in spontaneously aborted fetuses, demonstrating that almost all 45,X conceptions are lost to miscarriage. Indeed, the majority of live-born females with Turner's syndrome are diagnosed as mosaics, meaning that some proportion of their cells are 45,X while the rest are either 46,XX or 46,XY. The degree of clinical severity generally correlates inversely with the degree of mosaicism, so that females with a higher proportion of normal cells will tend to have a milder clinical outcome.

In contrast to Turner's syndrome, which results from the absence of a sex chromosome, three alternative conditions result from the presence of an extra sex chromosome: Klinefelter's syndrome, trisomy X, and 47,XYY syndrome. These conditions, each of which occurs in about 1 in 1,000 live births, are clinically mild, perhaps reflecting the fact that the Y chromosome carries relatively few genes, and, although the X chromosome is gene rich, most of these genes become transcriptionally silent in all but one X chromosome in each somatic cell (*i.e.*, all cells except eggs and sperm) via a process called X inactivation. The phenomenon of X inactivation prevents a female who carries two copies of the X chromosome in every cell from expressing twice the amount of gene products encoded exclusively on the X chromosome, in comparison with males, who carry a single X. In brief, at some point in early development, one X chromosome in each somatic cell of a female embryo undergoes chemical modification and is inactivated so that gene expression no longer occurs from that template. This process is apparently random in most embryonic tissues, so that roughly half of the cells in each somatic tissue will inactivate the maternal X, while the other half will inactivate the paternal X. Cells destined to give rise to eggs do not undergo X inactivation, and cells of the extra-embryonic tissues preferentially inactivate the paternal X, although the rationale for this preference is unclear. The inactivated X chromosome typically replicates later than other chromosomes, and it physically condenses to form a Barr body, a small structure found at the rim of the nucleus in female somatic cells between divisions (Figure 18). The discovery of X inactivation is generally attributed to British geneticist Mary Lyon, and it is therefore often called "lyonization." The result of X inactivation is that all normal females are mosaics with regard to this chromosome, meaning that they are composed of some cells that express genes only from the maternal X chromosome and others that express genes only from the paternal X chromosome. Although the process is apparently random, not every female has an exact 1 : 1 ratio of maternal to paternal X inactivation. Indeed, studies suggest that ratios of X inactivation can vary widely between different women. Finally, not all genes on the X chromosome are inactivated; a small number escape modification and remain actively expressed from both X chromosomes in the cell. Although this class of genes has not yet been fully characterized, aberrant expression of these genes has been

From the Cytogenetics Laboratory of Dr. Arthur Robinson, National Jewish Hospital and Research Center/National Asthma Center, Denver, Colorado

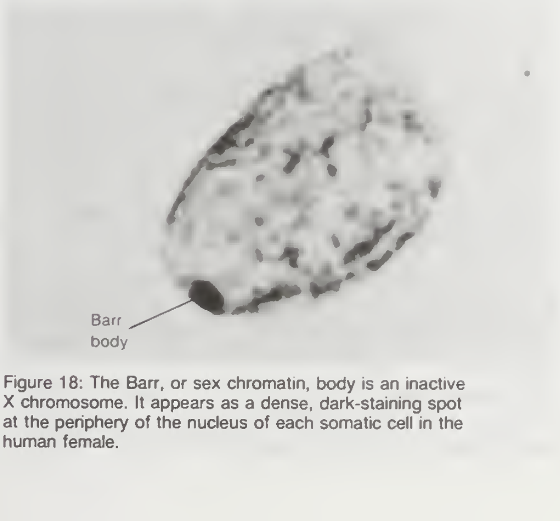


Figure 18: The Barr, or sex chromatin, body is an inactive X chromosome. It appears as a dense, dark-staining spot at the periphery of the nucleus of each somatic cell in the human female.

raised as one possible explanation for the phenotypic abnormalities experienced by individuals with too few or too many X chromosomes.

Klinefelter's syndrome

Klinefelter's syndrome (47,XXY) occurs in males and is associated with increased stature and infertility. Gynecomastia (*i.e.*, partial breast development in a male) is sometimes also seen. Males with Klinefelter's syndrome, like normal females, inactivate one of their two X chromosomes in each cell.

Trisomy X (47,XXX) is seen in females and is generally also considered clinically benign, although menstrual irregularities or sterility have been noted in some cases. Females with trisomy X inactivate two of the three X chromosomes in each of their cells.

Another condition, 47,XYY syndrome, also occurs in males and is associated with tall stature but few, if any, other clinical manifestations. There is some evidence of mild learning disability associated with each of the sex chromosome trisomies, although there is no evidence of mental retardation in these persons.

Finally, persons with karyotypes of 48,XXXY or 49,XXXXY have been reported but are extremely rare. These individuals show clinical outcomes similar to those seen in males with Klinefelter's syndrome but with slightly increased severity. In these persons the "$n - 1$ rule" for X inactivation still holds, so that all but one of the X chromosomes present in each somatic cell are inactivated.

Diseases associated with single-gene (Mendelian) inheritance. The term *Mendelian* is often used to denote patterns of genetic inheritance similar to those described for traits in the garden pea by Gregor Mendel in the 1860s. Disorders associated with single-gene Mendelian inheritance are typically categorized as autosomal dominant, autosomal recessive, or sex-linked. Each category is described briefly in this section.

Autosomal dominant inheritance. A disease trait that is inherited in an autosomal dominant manner can occur in either sex and can be transmitted by either parent. It manifests itself in the heterozygote (designated *Aa*), who receives a mutant gene (designated *A*) from one parent and a normal ("wild-type") gene (designated *a*) from the other. In such a case the pedigree (*i.e.*, a pictorial representation of family history) is vertical—that is, the disease passes from one generation to the next. Figure 19 illustrates the pedigree for a family with achondroplasia, an autosomal dominant disorder characterized by short-limbed dwarfism that results from a specific mutation in the fibroblast growth factor receptor 3 (*FGFR3*) gene. In pedigrees of this sort, circles refer to females and squares to males; two symbols directly joined at the midpoint represent a mating, and those suspended from a common overhead line represent siblings, with descending birth order from left to right. Solid symbols represent affected individuals, and open symbols represent unaffected individuals. The Roman numerals denote generations, whereas the Arabic numerals identify individuals within each generation. Each person listed in a pedigree may therefore be specified uniquely by a combination of one Roman and one Arabic numeral, such as II-1.

Achondroplasia

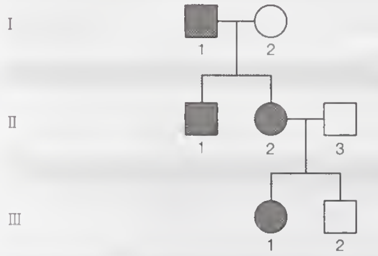


Figure 19: Pedigree of a family with a history of achondroplasia, an autosomal dominantly inherited disease (see text).

An individual who carries one copy of a dominant mutation (*Aa*) will producc two kinds of germ cells—eggs or sperm—typically in equal proportions; one half will bear the mutant gene (*A*), and the other will bear the normal gene (*a*). As a result, an affected heterozygote has a 50 percent chance of passing on the disease gene to each of his or her children. If an individual were to carry two copies of the dominant mutant gene (inherited from both parents), he or she would be homozygous (*AA*). The homozygote for a dominantly inherited abnormal gene may be equally affected with the heterozygote. Alternatively, he or she may be much more seriously affected; indeed, the homozygous condition may be lethal, sometimes even in utero or shortly after birth. Such is the case with achondroplasia, so that a couple with one affected partner and one unaffected partner will typically see half of their children affected, whereas a couple with both partners affected will see two-thirds of their surviving children affected and one-third unaffected, because 1 out of 4 conceptions will produce a homozygous fetus who will die before or shortly after birth.

Although autosomal dominant traits are typically evident in multiple generations of a family, they can also arise from new mutations, so that two unaffected parents, neither of whom carries the mutant gene in their somatic cells, can conceive an affected child. Indeed, for some disorders the new mutation rate is quite high; almost 7 out of 8 children with achondroplasia are born to two unaffected parents. Examples of autosomal dominant inheritance are common among human traits and diseases. More than 2,000 of these traits have been clearly identified; a sampling is given in Table 5.

Table 5: Human Conditions Attributable to a Single Dominant Gene

trait	conspicuous signs
Achondroplasia	dwarfism, large head, short extremities, short fingers and toes
Osteogenesis imperfecta	bone fragility, deafness
Huntington's disease	involuntary movement, emotional disturbance, dementia
Marfan's syndrome	long, thin extremities and fingers, eye and cardiovascular problems
Neurofibromatosis	pigmented spots (café au lait) on skin, skin tumours, occasional brain or other internal tumours

In many genetic diseases, including those that are autosomal dominant, specific mutations associated with the same disease present in different families may be uniform, such that every affected individual carries exactly the same molecular defect (allelic homogeneity), or they may be heterogeneous, such that tens or even hundreds of different mutations, all affecting the same gene, may be seen in the affected population (allelic heterogeneity). In some cases even mutations in different genes can lead to the same clinical disorder (genetic heterogeneity). Achondroplasia is characterized by allelic homogeneity, such that essentially all affected individuals carry exactly the same mutation.

With regard to the physical manifestations (*i.e.*, the phenotype) of some genetic disorders, a mutant gene may cause many different symptoms and may affect many different organ systems (pleiotropy). For example, along with the short-limbed dwarfism characteristic of achondroplasia, some individuals with this disorder also exhibit a long, narrow trunk, a large head with frontal bossing, and hyperextensibility of most joints, especially the knees. Similarly, for some genetic disorders, clinical severity may vary dramatically, even among affected members in the same family. These variations of phenotypic expression are called variable expressivity, and they are undoubtedly due to the modifying effects of either other genes or environmental factors. Finally, although for some disorders, such as achondroplasia, essentially all individuals carrying the mutant gene exhibit the disease phenotype, for other disorders some individuals who carry the mutant gene may express no apparent phenotypic abnormalities at all. Such unaffected individuals are called "non-penetrant," although they can pass on the mutant gene to their offspring, who could be affected.

Autosomal recessive inheritance. Nearly 2,000 traits have been related to single genes that are recessive; that is, their effects are masked by normal ("wild-type") dominant alleles and manifest themselves only in individuals homozygous for the mutant gene. A partial list of recessively inherited diseases is given in Table 6. For example, sickle cell anemia, a severe hemoglobin disorder, results only when a mutant gene (*a*) is inherited from both parents. Each of the latter is a carrier, a heterozygote with one normal gene and one mutant gene (*Aa*) who is phenotypically unaffected. The chance of such a couple producing a child with sickle cell anemia is one out of four for each pregnancy. For couples consisting of one carrier (*Aa*) and one affected individual (*aa*), the chance of their having an affected child is one out of two for each pregnancy.

Table 6: Human Characteristics Often Attributable to a Single Pair of Recessive Genes

trait	conspicuous signs
Albinism	lack of pigment in skin, hair, and eyes, with significant visual problems
Tay-Sachs disease	metabolic disorder that is lethal in childhood
Cystic fibrosis	chronic lung and intestinal symptoms
Phenylketonuria	enzyme disorder leading to mental retardation
Thalassemia	abnormal production of hemoglobin in red blood cells, leading to severe anemia

Phenylketonuria

Many autosomal recessive traits reflect mutations in key metabolic enzymes and result in a wide variety of disorders classified as inborn errors of metabolism. One of the best-known examples of this class of disorders is phenylketonuria (PKU), which results from mutations in the gene encoding the enzyme phenylalanine hydroxylase (PAH). PAH normally catalyzes the conversion of phenylalanine, an amino acid prevalent in dietary proteins and in the artificial sweetener aspartame, to another amino acid called tyrosine. In persons with PKU, dietary phenylalanine either accumulates in the body or some of it is converted to phenylpyruvic acid, a substance that normally is produced only in small quantities. Individuals with PKU tend to excrete large quantities of this acid, along with phenylalanine, in their urine. When infants accumulate high concentrations of phenylpyruvic acid and unconverted phenylalanine in their blood and other tissues, the consequence is mental retardation. Fortunately, with early detection, strict dietary restriction of phenylalanine, and supplementation of tyrosine, mental retardation can be prevented.

Since the recessive genes that cause inborn errors of metabolism are individually rare in the gene pool, it is not often that both parents are carriers; hence, the diseases are relatively uncommon. If the parents are related (consanguineous), however, they will be more likely to have inherited the same mutant gene from a common ancestor. For this reason, consanguinity is often more common in the parents of those with rare, recessive inherited diseases. The pedigree of a family in which PKU has occurred follows (see Figure 20). This pedigree demonstrates that the affected individuals for recessive diseases are usually siblings in one generation—the pedigree tends to be "horizontal" rather than "vertical" as in dominant inheritance.

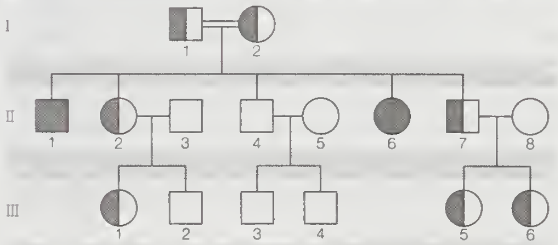


Figure 20: Pedigree of a family in which the gene for phenylketonuria is segregating. The half-solid symbols represent heterozygous carriers of the trait; the double line between the I-1 and I-2 signifies a consanguineous mating.

Sex-linked inheritance. In humans, there are hundreds of genes located on the X chromosome that have no counterpart on the Y chromosome. The traits governed by these genes thus show sex-linked inheritance. This type of inheritance has certain unique characteristics, which include the following: (1) There is no male-to-male (father-to-son) transmission, since sons will, by definition, inherit the Y rather than the X chromosome. (2) The carrier female (heterozygote) has a 50 percent chance of passing the mutant gene to each of her children; sons who inherit the mutant gene will be hemizygotes and will manifest the trait, while daughters who receive the mutant gene will be unaffected carriers. (3) Males with the trait will pass the gene on to all of their daughters, who will be carriers. (4) Most sex-linked traits are recessively inherited, so that heterozygous females generally do not display the trait. Table 7 lists some sex-linked conditions. Figure 21 shows a pedigree of a family in which a mutant gene for hemophilia A, a sex-linked recessive disease, is segregating. Hemophilia A gained notoriety in early studies of human genetics because it affected at least 10 males among the descendants of Queen Victoria, who was a carrier.

Hemophilia A

Table 7: Some Sex-Linked Recessively Inherited Conditions

trait	conspicuous signs
Hemophilia A	bleeding tendency with joint involvement
Duchenne's muscular dystrophy	progressive muscle weakness
Lesch-Nyhan syndrome	cerebral palsy, self-mutilation
Fragile-X syndrome	mental retardation, characteristic facies

Hemophilia A, the most widespread form of hemophilia, results from a mutation in the gene encoding clotting factor VIII. Because of this mutation, affected males cannot produce functional factor VIII, so that their blood fails to clot properly, leading to significant and potentially life-threatening loss of blood after even minor injuries. Bleeding into joints commonly occurs as well and may be crippling. Therapy consists of avoiding trauma and of administering injections of purified factor VIII, which was once isolated from outdated human blood donations but can now be made in large amounts through recombinant DNA technology.

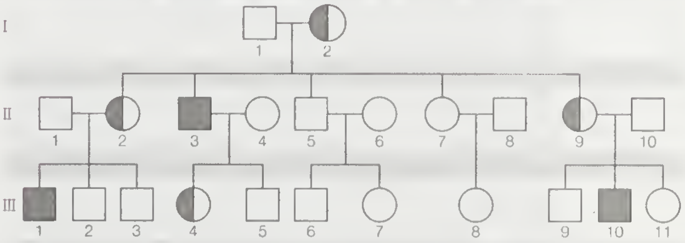


Figure 21: Pedigree of a family with a history of hemophilia A, a sex-linked recessively inherited disease.The half-solid circles represent female carriers (heterozygotes); the solid squares signify affected males (hemizygotes).

Although heterozygous female carriers of X-linked recessive mutations generally do not exhibit traits characteristic of the disorder, cases of mild or partial phenotypic expression in female carriers have been reported, resulting from non-random X inactivation.

Diseases associated with single-gene (non-Mendelian) inheritance. Although disorders resulting from single gene defects that demonstrate Mendelian inheritance are perhaps better understood, it is now clear that a significant number of single-gene diseases also exhibit distinctly non-Mendelian patterns of inheritance. Among these are such disorders that result from triplet repeat expansions within or near specific genes (*e.g.,* Huntington's disease and fragile-X syndrome); a collection of neurodegenerative disorders, such as Leber hereditary optic neuropathy (LHON), that result from inherited mutations in the mitochondrial

DNA; and diseases that result from mutations in imprinted genes (*e.g.,* Angelman's syndrome and Prader-Willi syndrome).

Triplet repeat expansions. At least a dozen different disorders are now known to result from triplet repeat expansions in the human genome, and these fall into two groups: (1) those that involve a polyglutamine tract within the encoded protein product that becomes longer upon expansion of a triplet repeat, an example of which is Huntington's disease; and (2) those that have unstable triplet repeats in non-coding portions of the gene that, upon expansion, interfere with appropriate expression of the gene product, an example of which is fragile-X syndrome (see Figure 22). Both groups of disorders exhibit a distinctive pattern of non-Mendelian inheritance termed anticipation, in which, following the initial appearance of the disorder in a given family, subsequent generations tend to show both increasing frequency and increasing severity of the disorder. This phenotypic anticipation is paralleled by increases in the relevant repeat length as it is passed from one generation to the next, with increasing size leading to increasing instability, until a "full expansion" mutation is achieved, generally several generations following the initial appearance of the disorder in the family. The full expansion mutation is then passed to subsequent generations in a standard Mendelian fashion—for example, autosomal dominant for Huntington's disease and sex-linked for fragile-X syndrome.

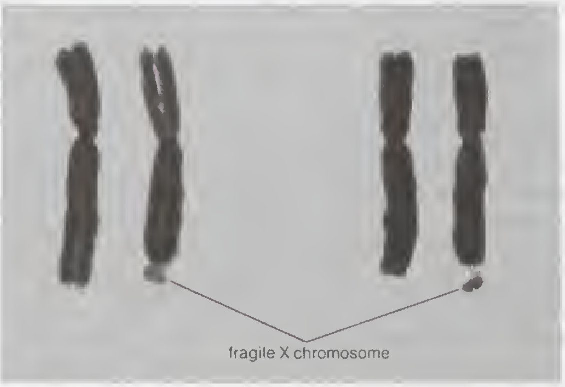


Figure 22: *The fragile X chromosome.*
The right-hand member in each of these two pairs of X chromosomes is a fragile X; the leader points to the fragile site at the tip of the long arm. Males hemizygous for this chromosome exhibit the fragile-X syndrome of mild to moderate mental retardation.

Mitochondrial DNA mutations. Disorders resulting from mutations in the mitochondrial genome demonstrate an alternative form of non-Mendelian inheritance, termed maternal inheritance, in which the mutation and disorder are passed from mothers—never from fathers—to all of their children. The mutations generally affect the function of the mitochondrion, compromising, among other processes, the production of cellular adenosine triphosphate (ATP). Severity and even penetrance can vary widely for disorders resulting from mutations in the mitochondrial DNA, generally believed to reflect the combined effects of heteroplasmy (*i.e.,* mixed populations of both normal and mutant mitochondrial DNA in a single cell) and other confounding genetic or environmental factors. There are close to 50 mitochondrial genetic diseases currently known.

Imprinted gene mutations. Finally, some genetic disorders are now known to result from mutations in imprinted genes. Genetic imprinting involves a sex-specific process of chemical modification to the imprinted genes, so that they are expressed unequally, depending on the sex of the parent of origin. So-called maternally imprinted genes are generally expressed only when inherited from the father, and so-called paternally imprinted genes are generally expressed only when inherited from the mother. The disease gene associated with Prader-Willi syndrome is maternally imprinted, so that although every child inherits two copies of the gene (one maternal, one paternal), only the paternal copy is expressed. If the paternally inherited copy carries a mutation, the child will be left with no functional copies of the gene expressed, and the clinical traits of Prader-Willi syndrome will result. Similarly, the disease gene associated with Angelman's syndrome is paternally imprinted, so that although every child inherits two copies of the gene, only the maternal copy is expressed. If the maternally inherited copy carries a mutation, the child again will be left with no functional copies of the gene expressed, and the clinical traits of Angelman's syndrome will result. Individuals who carry the mutation but received it from the "non-imprinting" parent can certainly pass it on to their children, although they will not exhibit clinical features of the disorder.

Upon rare occasion, persons are identified with an imprinted gene disorder who show no family history and who do not appear to carry any mutation in the expected gene. These cases are now known to result from uniparental disomy, a phenomenon whereby a child is conceived who carries the normal complement of chromosomes, but who has inherited both copies of a given chromosome from the same parent, rather than one from each parent, as is the normal fashion. If any key genes on that chromosome are imprinted in the parent of origin, the child may end up with no expressed copies, and a genetic disorder may result. Similarly, other genes may be overexpressed in cases of uniparental disomy, perhaps also leading to clinical complications. Finally, uniparental disomy can account for very rare instances whereby two parents, only one of whom is a carrier of an autosomal recessive mutation, can nonetheless have an affected child, in the circumstance that the child inherits two mutant copies from the carrier parent.

Diseases caused by multifactorial inheritance. Genetic disorders that are multifactorial in origin represent probably the single largest class of inherited disorders affecting the human population. By definition, these disorders involve the influence of multiple genes, generally acting in concert with environmental factors. Such common conditions as cancer, heart disease, and diabetes are now considered to be multifactorial disorders. Indeed, improvements in the tools used to study this class of disorders have enabled the assignment of specific contributing gene loci to a number of common traits and disorders. Identification and characterization of these contributing genetic factors may not only enable improved diagnostic and prognostic indicators but may also identify potential targets for future therapeutic intervention.

Table 8: Common Multifactorially Determined Diseases
Alcoholism
Alzheimer's disease
Cancer
Coronary heart disease
Diabetes
Epilepsy
Hypertension
Obesity
Schizophrenia

Table 8 lists some conditions associated with multifactorial inheritance. Because the genetic and environmental factors that underlie these conditions are often unknown, the risks of recurrence are usually arrived at empirically. In general, it can be said that risks of recurrence are not as great for multifactorial conditions as for single-gene diseases and that the risks vary with the number of relatives affected and the closeness of their relationship. Moreover, close relatives of more severely affected individuals (*e.g.,* those with bilateral cleft lip and cleft palate) are generally at greater risk than those related to persons with a less severe form of the same condition (*e.g.,* unilateral cleft lip).

GENETICS OF CANCER

Although at least 90 percent of all cancers are sporadic, meaning that they do not seem to run in families, nearly 10 percent of cancers are now recognized as familial, and some are actually inherited in an apparently autosomal dominant manner. Cancer may therefore be considered a multifactorial disease, resulting from the combined influence of many genetic factors acting in concert with environmental insults (*e.g.,* ultraviolet radiation, cigarette smoke, and viruses).

Cancers, both familial and sporadic, generally arise from alterations in one or more of three classes of genes: oncogenes, tumour suppressor genes, and genes whose products participate in genome surveillance—for example, in DNA damage repair. All of these functions are described in the article CANCER. For familial cancers, affected members inherit one mutant copy of a gene that falls into one of the latter two classes. That mutation alone is not sufficient to cause cancer, but it predisposes individuals to the disease because they are now either more sensitive to spontaneous somatic mutations, as in the case of altered tumour suppressor genes, or are more prone to experience mutations, as in the case of impaired DNA repair enzymes. Of course, sporadic cancers can also arise from mutations in these same classes of genes, but because all of the mutations must arise in the individual *de novo,* as opposed to being inherited, they generally appear only later in life, and they do not run in families.

Retinoblastoma

Retinoblastoma, an aggressive tumour of the eye that typically occurs in childhood, offers perhaps one of the clearest examples of the interplay between inherited and somatic mutations in the genesis of cancer. Current data suggest that 60 to 70 percent of all cases of retinoblastoma are sporadic, while the rest are inherited. The relevant gene, *RB,* encodes a protein that normally functions as a suppressor of cell cycle progression and is considered a classic tumour suppressor gene. Children who inherit one mutant copy of the *RB* gene are at nearly 100 percent risk to develop retinoblastoma because the probability that their one remaining functional *RB* gene will sustain a mutation in at least one retinal cell is nearly assured. In contrast, children who inherit two functional copies of the *RB* gene must experience two mutations at the *RB* locus in the same retinal cell in order to develop retinoblastoma; this is a very rare event. This "two-hit" hypothesis of retinoblastoma formation, originated by Alfred Knudson, Jr., has provided a foundation upon which most subsequent theories of the genetic origins of familial cancer have been built.

Recent studies of both breast and colon cancers have revealed that, like retinoblastoma, these cancers are predominantly sporadic, although a small proportion are clearly familial. Sporadic breast cancer generally appears late in life, while the familial forms can present much earlier, often before age 40. For familial breast cancer, inherited mutations in one of two specific genes, *BRCA1* and *BRCA2,* account for at least half of the cases observed. The *BRCA1* and *BRCA2* genes both encode protein products believed to function in the pathways responsible for sensing and responding to DNA damage in cells. While a woman in the general population has about a 10 percent lifetime risk of developing breast cancer, half of all women with *BRCA1* or *BRCA2* mutations will develop breast cancer by age 50, and close to 90 percent will develop the disease by age 80. Women with *BRCA1* mutations are also at increased risk to develop ovarian tumours. As with retinoblastoma, both men and women who carry *BRCA1* or *BRCA2* mutations, whether they are personally affected or not, can pass the mutated gene to their offspring, although carrier daughters are much more likely than carrier sons to develop breast cancer.

BRCA1 and *BRCA2* genes

Two forms of familial colon cancer, hereditary nonpolyposis colorectal cancer (HNPCC) and familial adenomatous polyposis (FAP), have also been linked to predisposing mutations in specific genes. Persons with familial HNPCC have inherited mutations in one or more of their DNA mismatch repair genes, predominantly *MSH2* or *MLH1.* Similarly, persons with FAP carry inherited mutations in their *APC* genes, the protein product of which normally functions as a tumour suppressor. For persons in both categories, the combination of inherited and somatic mutations results in a nearly 100 percent lifetime risk of developing colon cancer.

Although most cancer cases are not familial, all are undoubtedly diseases of the genetic material of somatic cells. Studies of large numbers of both familial and sporadic cancers have led to the conclusion that cancer is a disease of successive mutations, acting in concert to deregulate normal cell growth, provide appropriate blood supply to the growing tumour, and ultimately enable tumour cell movement beyond normal tissue boundaries to achieve metastasis (*i.e.,* the dissemination of cancer cells to other parts of the body).

Many of the agents that cause cancer (*e.g.,* X rays, certain chemicals) also cause mutations or chromosome abnormalities. For example, a large fraction of sporadic tumours have been found to carry oncogenes, altered forms of normal genes (proto-oncogenes) that have sustained a somatic "gain-of-function" mutation. An oncogene may be carried by a virus, or it can result from a chromosomal rearrangement, as is the case in chronic myelogenous leukemia, a cancer of the white blood cells characterized by the presence of the so-called Philadelphia chromosome in affected cells. The Philadelphia chromosome arises from a translocation in which one half of the long arm of chromosome 22 becomes attached to the end of the long arm of chromosome 9, creating the dominant oncogene *BCR/abl* at the junction point. The specific function of the *BCR/abl* fusion protein is not entirely clear. Another example is Burkitt's lymphoma, in which a rearrangement between chromosomes places the *myc* gene from chromosome 8 under the influence of regulatory sequences that normally control expression of immunoglobulin genes. This deregulation of *myc,* a protein involved in mediating cell cycle progression, is thought to be one of the major steps in the formation of Burkitt's lymphoma.

COGNITIVE AND BEHAVIORAL GENETICS

Mental activities, expressed in human behaviour, are intimately related to physical activities in the brain and nervous system. In 1929 British physician Sir Archibald Garrod emphasized this when he wrote: "Each one of us differs from his fellows, not only in bodily structure and the proteins which enter into his composite, but apart from, or rather in consequence of, such individualities, men differ in mental outlook, character and ability." Since that time, many investigators have sought to analyze the molecular and cellular components of behaviour in order to relate genes to such abstractions as intellect, temperament, and the emotions. Because the brain is ultimately responsible for behavioral development, neurobiologists have attempted to understand the unusual precision by which this organ's various regions are interconnected and the intricate chemical signals that, under genetic control, organize its complicated nerve fibre circuits.

"Nature" versus "nurture"

Some of the most powerful experiments to dissect the "nature"-versus-"nurture" aspects of human intelligence and behaviour have involved studies of twins, both monozygotic (identical) and dizygotic (fraternal). Cognitive or behavioral characteristics that are entirely under genetic control would be predicted to be the same, or concordant, in monozygotic twins, who share identical genes regardless of their upbringing. These same characteristics would be predicted to be less concordant in dizygotic twins, who share, on average, only half of their genes. Comparison of the concordance rates among monozygotic and dizygotic twins monitored for different traits allows an estimate of the heritability of those traits—that is, the proportion of population variation for a given trait that can be attributed to genes. A heritability value of 1.0 implies a purely genetic basis for a trait, and a value of 0.0 implies a purely environmental basis. Intelligence, measured as IQ, has a heritability value of 0.5, indicating that both genetics and the environment play major roles. In contrast, schizophrenia has a value of 0.7, and both autism and bipolar disorder have heritability values of 1.0. Clearly, genetics play a large role in determining not only how our bodies look and function but also how we think and feel.

GENETIC DAMAGE FROM ENVIRONMENTAL AGENTS

We are exposed to many agents, both natural and manmade, that can cause genetic damage. Among these agents are viruses; compounds produced by plants, fungi, and bacteria; industrial chemicals; products of combustion; alcohol; ultraviolet and ionizing radiation; and even the oxygen that we breathe. Many of these agents have long been unavoidable, and consequently human beings have evolved defenses to minimize the damage that they cause and ways to repair the damage that cannot be avoided.

Viruses. Viruses survive by injecting their genetic material into living cells with the consequence that the biochemical machinery of the host cell is subverted from serving its own needs to serving the needs of the virus. During this process the viral genome often integrates itself into the genome of the host cell. This integration, or insertion, can occur either in the intergenic regions that make up the vast majority of human genomes, or it can occur in the middle of an important regulatory sequence or even in the region coding for a protein, *i.e.,* a gene. In either of the latter two scenarios, the regulation or function of the interrupted gene is lost. If that gene encodes a protein that normally regulates cell division, the result may be unregulated cell growth and division. Alternatively, some viruses carry dominant oncogenes in their genomes, which can transform an infected cell and start it on the path toward cancer.

Finally, viruses can cause mutations leading to cancer by killing the infected cell. Indeed, one of the body's defenses against viral infection involves recognizing and killing infected cells. The death of cells necessitates their replacement by the division of uninfected cells, and the more cell division that occurs, the greater the likelihood of a mutation arising from the small but finite infidelity in DNA replication. Among the viruses that can cause cancer are: Epstein-Barr virus, papillomaviruses, hepatitis B and C viruses, retroviruses (*e.g.,* human immunodeficiency virus), and herpesviruses.

Plants, fungi, and bacteria. During the ongoing struggle for survival, organisms have evolved toxic compounds as protection against predators or simply to gain competitive advantage. At the same time, these organisms have evolved mechanisms that make themselves immune to the effects of the toxins that they produce. Plants in particular utilize this strategy since they are rooted in place and cannot escape from predators. One-third of the dry weight of some plants can be accounted for by the toxic compounds that are collectively referred to as alkaloids. *Aspergillus flavus,* a fungus that grows on stored grain and peanuts, produces a powerful carcinogen called aflatoxin that can cause liver cancer. Bacteria produce many proteins that are toxic to the infected host such as diphtheria toxin. They also produce proteins called bacteriocins that are toxic to other bacteria. Toxins can cause mutations indirectly by causing cell death, which necessitates replacement by cell division, thus enhancing the opportunity for mutation. Cyanobacteria that grow in illuminated surface water produce several carcinogens such as microcystin, saxitoxin, and cylindrospermopsin that can also cause liver cancer.

Industrial chemicals. Tens of thousands of different chemicals are routinely used in the production of plastics, fuels, food additives, and even medicines. Many of these chemicals are mutagens, and some have been found to be carcinogenic (cancer producing) in rats or mice. A relatively easy and inexpensive test for mutagenicity, the Ames test, utilizes mutant strains of the bacterium *Salmonella typhimurium* and can be completed in a few days. Testing for carcinogenesis, on the other hand, is very time consuming and expensive because the test substance must be administered to large numbers of laboratory animals, usually mice, for months before the tissues can be examined for cancers. For this reason, the number of known mutagens far exceeds the number of known carcinogens. Furthermore, animal tests for carcinogenesis are not completely predictive of the effects of the test chemical on humans for several reasons. First, the abilities of laboratory animals and humans to metabolize and excrete specific chemicals can differ greatly. In addition, in order to avoid the need to test each chemical at a range of doses, each chemical is usually administered at the maximum tolerated dose. At such high doses, toxicity and cell death occurs, necessitating cell replacement by growth and cell division; cell division, in turn, increases the opportunity for mutation and hence for cancer. Alternatively, unusually high doses of a chemical may actually mask the carcinogenic potential of a compound because damaged cells may die rather than survive in mutated form.

Combustion products. The burning of fossil fuels quite literally powers modern industrial societies. If the combustion of such fuels were complete, the products would be carbon dioxide and water. However, combustion is rarely complete, as is evidenced by the visible smoke issuing from chimneys and from the exhausts of diesel engines. Moreover, in addition to the particulates that we can see, incomplete combustion produces a witch's brew of volatile compounds that we do not see; and some of these, such as the dibenzodioxins, are intensely mutagenic and have been demonstrated to cause cancer in laboratory rodents. Epidemiological data indicates that dioxins are associated with increased risk of a variety of human cancers. The health consequences of combustion are further increased by impurities in fossil fuels and in the oxygen that supports their burning. For example, coal contains sulfur, mercury, lead, and other elements in addition to carbon. During combustion sulfur becomes sulfur dioxide, and that, in turn, gives rise to sulfurous and sulfuric acids. The mercury in the fuel is emitted as a vapour that is very toxic. Atmospheric nitrogen is oxidized at the high temperature of combustion.

Cigarette smoke

The smoke from a cigarette, drawn directly into the lungs, imparts a large number of particulates, as well as a host of volatile compounds, directly into the airways and alveoli. Some of the volatile compounds are toxic in their own right and others, such as hydroquinones, slowly oxidize, producing genotoxic free radicals. As macrophages in the lungs attempt to engulf and eliminate particulates, they cause the production of mutagenic substances. A large fraction of lung cancers are attributable to cigarette smoking, which is also a risk factor for atherosclerosis, high blood pressure, heart attacks, and strokes.

Alcohol. Moderate consumption of alcohol (ethanol) is well tolerated and may even increase life span. However, alcohol is a potentially toxic substance and one of its metabolites, acetaldehyde, is a mild mutagen. Hence, it is not surprising that the chronic consumption of alcohol leads to liver cirrhosis and other untoward effects. Consumption of alcohol during pregnancy can cause fetal alcohol syndrome, which is characterized by low birth weight, mental retardation, and congenital heart disease.

Skin cancer and the ozone layer

Ultraviolet radiation. Due to human activities that result in the release of volatile halocarbon compounds, such as the refrigerant freon and the solvent carbon tetrachloride, the chlorine content of the upper atmosphere is increasing. Chlorine catalyzes the decomposition of ozone, which shields the Earth from ultraviolet radiation that is emitted from the Sun. The Earth's ozone shield has been progressively depleted, most markedly over the polar regions but also measurably so over the densely populated regions of northern Europe, Australia, and New Zealand. One consequence has been an increase in a variety of skin cancers, including melanoma, in those areas. Steps have been taken to stop the release of halocarbons, but the depletion of the ozone layer will nonetheless persist and may worsen for at least several decades.

Ultraviolet light, when acting on DNA, can lead to covalent linking of adjacent pyrimidine bases. Such pyrimidine dimerization is mutagenic, but this damage can be repaired by an enzyme called photolyase, which utilizes the energy of longer wavelengths of light to cleave the dimers. However, people with a defect in the gene coding for photolyase develop xeroderma pigmentosum, a condition characterized by extreme sensitivity to sunlight. These individuals develop multiple skin cancers on all areas of exposed skin, such as the head, neck, and arms.

Ultraviolet light can also be damaging because of photosensitization, the facilitation of photochemical processes. One way that photosensitizers work is by absorbing a photon and then transferring the energy inherent in that photon to molecular oxygen, thus converting the less active ground-state molecular oxygen into a very reactive excited state, referred to as singlet oxygen, that can attack a variety of cellular compounds, including DNA. Diseases that have a photosensitizing component include lupus and porphyrias. In addition to photosensitizers that occur naturally in the human body, some foods and medicines (*e.g.*, tetracyclines and neuroleptics) also act in this way, producing painful inflammation and blistering of the skin following even modest exposure to the sun.

Ionizing radiation. X rays and gamma rays are sufficiently energetic to cleave water into hydrogen atoms and hydroxyl radicals and are consequently referred to as ionizing radiation. Ionizing radiation and the products of the cleavage of water are able to damage all biological macromolecules, including DNA, proteins, and polysaccharides, and they have long been recognized as being mutagenic, carcinogenic, and lethal. People are routinely exposed to natural sources of ionizing radiation, such as cosmic rays and radioisotopes such as carbon-14 and radon. They are also exposed to X rays and man-made radioisotopes used for diagnostic purposes, and some people have been exposed to radioactive fallout from nuclear weapon tests and reactor accidents. Such exposures would be much more damaging were it not for multiple mechanisms of DNA repair, which have evolved to deal with simple errors in replication as well as damage from naturally occurring sources of damage.

Molecular oxygen. Molecular oxygen (O_2), although essential for life, must be counted among the environmental toxins and mutagens. Because of its unusual electronic structure, O_2 is most easily reduced not by electron pairs but rather by single electrons added one at a time. As O_2 is converted into water, superoxide (O_2^-), hydrogen peroxide (H_2O_2), and a hydroxyl radical (HO) are produced as intermediates. O_2^- can initiate free radical oxidation of important metabolites, inactivate certain enzymes, and cause release of iron from specific enzymes. The second intermediate, H_2O_2, is a strong oxidant and can give rise to an even more potent oxidant, namely HO, when it reacts with ferrous iron. Thus, O_2^- and H_2O_2 can collaborate in the formation of the destructive HO and can subsequently lead to DNA damage, mutagenesis, and cell death. Breathing 100 percent oxygen causes damage to the alveoli, which leads to accumulation of fluid in the lungs. Thus, paradoxically, prolonged exposure to hyperoxia causes death due to lack of oxygen.

Humans have evolved multiple defense systems to counter the toxicity and mutagenicity of O_2. Thus, O_2 is rapidly converted into O_2 and H_2O_2 by a family of enzymes called superoxide dismutases. H_2O_2, in turn, is eliminated by other enzymes called catalases and peroxidases, which convert it into O_2 and water.

A few genetic diseases are known to be related to oxygen radicals or to the enzymes that defend against them. Chronic granulomatous disease (CGD) is caused by a defect in the ability of the phagocytic leukocytes to mount the respiratory burst, part of the body's defense against infection. Upon contacting microorganisms and engulfing them, phagocytes greatly increase their consumption of O_2 (the respiratory burst) while releasing O_2^-, H_2O_2, hypochlorite (HOCl), and other agents that kill the microbe. The reduction of O_2 to O_2^- is caused by a multicomponent enzyme called the NADPH oxidase. A defect in any of the components of this oxidase will lead to the absence of the respiratory burst, giving rise to the constant infections indicative of CGD. Before the discovery and clinical application of antibiotics, people born with CGD died from infection during early childhood.

Lou
Gehrig's
disease

Another such genetic disease is the familial form of amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease, which is characterized by late-onset progressive paralysis due to the loss of motor neurons. Approximately 20 percent of cases of ALS have been shown to result from mutations affecting the enzyme superoxide dismutase. The disease is genetically dominant, so that the mutant enzyme causes the disease even when half of the superoxide dismutase present in cells exists in the normal form. Interestingly, most of the mutant variants retain full catalytic activity.

MANAGEMENT OF GENETIC DISEASE

The management of genetic disease can be divided into counseling, diagnosis, and treatment. In brief, the fundamental purpose of genetic counseling is to help the individual or family understand their risks and options and to empower them to make informed decisions. Diagnosis of genetic disease is sometimes clinical, based on the presence of a given set of symptoms, and sometimes molecular, based on the presence of a recognized gene mutation, whether clinical symptoms are present or not. The cooperation of family members may be required to achieve diagnosis for a given individual, and, once accurate diagnosis of that individual has been determined, there may be implications for the diagnoses of other family members. Balancing privacy issues within a family with the ethical need to inform individuals who are at risk for a particular genetic disease can become extremely complex.

Finally, although effective treatments exist for some genetic diseases, for others there are none. It is perhaps this latter set of disorders that raises the most troubling questions with regard to presymptomatic testing, because phenotypically healthy individuals can be put in the position of hearing that they are going to become ill and potentially die and that there is nothing they or anyone else can do to stop it. Fortunately, with time and research, this set of disorders is slowly becoming smaller.

Genetic counseling. Genetic counseling represents the most direct medical application of the advances in understanding of basic genetic mechanisms. Its chief purpose is to help people make responsible and informed decisions concerning their own health or that of their children. Genetic counseling, at least in democratic societies, is non-directive; the counselor provides information, but decisions are left up to the individual or the family.

Calculating risks of known carriers. Most couples who present themselves for pre-conceptional counseling fall into one of two categories: those who have already had a child with genetically based problems, and those who have one or more relatives with a disease they think might be inherited. The counselor must confirm the diagnosis in the affected person with meticulous accuracy, so as to rule out the possibility of alternative explanations for the clinical symptoms observed. A careful family history permits construction of a pedigree that may illuminate the nature of the inheritance (if any), may affect the calculation of risk figures, and may bring to light other genetic influences. The counselor, a certified health-care professional with special training in medical genetics, must then decide whether the disease in question has a strong genetic component and, if so, whether the heredity is single-gene, chromosomal, or multifactorial.

In the case of single-gene Mendelian inheritance, the disease may be passed on as an autosomal recessive, autosomal dominant, or sex-linked recessive trait, as discussed in the section *Classes of genetic disease.* If the prospective parents already have a child with an autosomal recessive inherited disease, they both are considered by definition to be carriers, and there is a 25 percent risk that each future child will be affected. If one of the parents carries a mutation known to cause an autosomal dominant inherited disease, whether that parent is clinically affected or not, there is a 50 percent risk that each future child will inherit the mutation and therefore may be affected. If, however, the couple has borne a child with an autosomal dominant inherited disease though neither parent carries the mutation, then it will be presumed that a spontaneous mutation has occurred and that there is not a markedly increased risk for recurrence of the disease in future children. There is a caveat to this reasoning, however, because there is also the possibility that the new mutation might have occurred in a progenitor germ cell in one of the parents, so that some unknown proportion of that individual's eggs or sperm may carry the mutation, even though it is absent from the somatic cells—including blood, which is generally the tissue sampled for testing. This scenario is called germ-line mosaicism. With regard to X-linked disorders, if the pedigree or carrier testing suggests that the mother carries a gene for a sex-linked disease, there is a 50 percent chance that each son will be affected and that each daughter will be a carrier.

Counseling for chromosomal inheritance most frequently involves either an inquiring couple (consultands) who have had a child with a known chromosomal disorder, such as Down syndrome, or a couple who have experienced multiple miscarriages. To provide the most accurate recurrence risk values to such couples, both parents should be karyotyped to determine if one may be a balanced translocation carrier. Balanced translocations refer to genomic rearrangements in which there is an abnormal covalent arrangement of chromosome segments, although there is no net gain or loss of key genetic material. If both parents exhibit completely normal karyotypes, the recurrence risks cited are low and are strictly empirical.

Most of the common hereditary birth defects, however, are multifactorial. (See above *Diseases caused by multifactorial inheritance.*) If the consulting couple have had one affected child, the empirical risk for each future child will be about 3 percent. If they have borne two affected children, the chance of recurrence will rise to about 10 percent. Clearly these are population estimates, so that the risks within individual families may vary.

Estimating probability: Bayes's theorem. As described above, the calculation of risks is relatively straightforward when the consultands are known carriers of diseases due to single genes of major effect that show regular Mendelian inheritance. For a variety of reasons, however, the parental genotypes frequently are not clear and must be approximated from the available family data. Bayes's theorem, a statistical method first devised by the English clergyman-scientist Thomas Bayes in 1763, can be used to assess the relative probability of two or more alternative possibilities (*e.g.,* whether a consultand is or is not a carrier). The likelihood derived from the appropriate Mendelian law (prior probability) is combined with any additional information that has been obtained from the consultand's family history or from any tests performed (conditional probability). A joint probability is then determined for each alternative outcome by multiplying the prior probability by all conditional probabilities. By dividing the joint probability of each alternative by the sum of all joint probabilities, the posterior probability is arrived at. Posterior probability is the likelihood that the individual, whose genotype is uncertain, either carries the mutant gene or does not.

This method can be applied to, for example, the sex-linked recessive disease Duchenne's muscular dystrophy (DMD). In this example, the consultand wishes to know her risk of having a child with DMD. The family's pedigree is illustrated in Figure 23. It is known that the consultand's grandmother (I-2) is a carrier, since she had two affected sons (spontaneous mutations occurring in both brothers would be extremely unlikely). What is uncertain is whether the consultand's mother (II-4) is also a carrier. The Bayesian method for calculating the consultand's risk is as follows:

	likelihood that consultand's mother (II-4) is a carrier	likelihood that consultand's mother (II-4) is not a carrier
prior probability	1/2	1/2
conditional probability	1/4 (for each of her sons there was a 1/2 chance of being unaffected)	1
joint probability	$1/2 \times 1/4 = 1/8$	$1 \times 1/2 = 1/2$
posterior probability	$\frac{1/8}{1/8 + 1/2} = 1/5$	$\frac{1/2}{1/8 + 1/2} = 4/5$

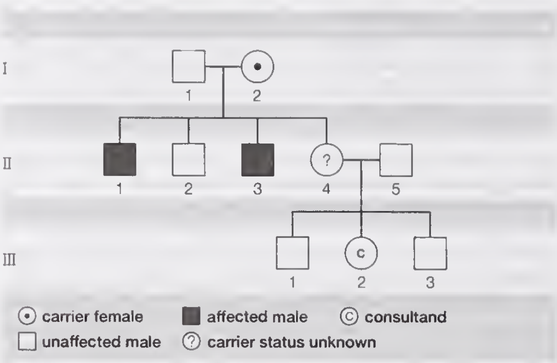


Figure 23: Pedigree of a family with a history of Duchenne's muscular dystrophy. The consultand (III-2) wishes to know her risk of having an affected child.

If II-4 is a carrier (risk = 1/5), then there is a 1/2 chance that the consultand is also a carrier, so her total empirical risk is 1/5 × 1/2 = 1/10. If she becomes pregnant, there is a 1/2 chance that her child will be male and a 1/2 chance that the child, regardless of sex, will inherit the familial mutation. Hence, the total empirical risk for the consultand (III-2) to have an affected child is 1/10 × 1/2 × 1/2 = 1/40. Of course, if the familial mutation is known, presumably from molecular testing of an affected family member, the carrier status of III-2 could be determined directly by molecular analysis, rather than estimated by Bayesian calculation. If the family is cooperative and an affected member is available for study, this is clearly the most informative route to follow, because the risk for the consultand to carry the familial mutation would be either 1 or 0, and not 1/10. If her risk is 1, then each of her sons will have a 1/2 chance of being affected. If her risk is 0, none of her children will be affected (unless a new mutation occurs, which is very rare).

Options available

After determining the nature of the heredity, the counselor discusses with the consultand the likely risks and the available options to minimize impact of those risks on the individual and the family. In the case of a couple in which one member has a family history of a genetic disorder—for example, cystic fibrosis—typical options might include any of the following choices: (1) Accept the risks and take a chance that any future children may be affected. (2) Seek molecular testing for known mutations for cystic fibrosis in relevant family members to determine with greater accuracy whether either or both prospective parents are carriers for this recessive disorder. (3) If both members of the couple are determined to be carriers, utilize donor sperm for artificial insemination. This option is a good genetic solution only if the husband carries a dominant mutation or if both parents are carriers of a recessive mutation. If the recessive trait is reasonably common, as are mutations for cystic fibrosis, however, it would be reasonable to ask that the sperm donor be checked for carrier status before pursuing this option. (4) Proceed with natural reproduction, but pursue prenatal diagnosis with the possibility of selective termination of an affected pregnancy, if desired by the parents. (5) Pursue in vitro fertilization with donor eggs, if the woman is the at-risk partner, or use both eggs and sperm from the couple but employ pre-implantation diagnostics to select only unaffected embryos for implantation (see below). (6) Decide against biological reproduction because the risks and available options are unacceptable; possibly pursue adoption.

Diagnosis. *Prenatal diagnosis.* Perhaps one of the most sensitive areas of medical genetics is prenatal diagnosis, the genetic testing of an unborn fetus, because of fears of eugenic misuse or because some couples may choose to terminate a pregnancy depending on the outcome of the test. Nonetheless, prenatal testing in one form or another is now almost ubiquitous in most industrialized nations, and recent advances both in testing technologies and in the set of "risk factor" genes to be screened promise to make prenatal diagnosis even more widespread. Indeed, parents may soon be able to ascertain not only information about the sex and health status of their unborn child but also about his or her complexion, personality, and intellect. Whether parents should have access to all of this information and how they may choose to use it is a matter of much debate.

Current forms of prenatal diagnosis can be divided into two classes, those that are apparently noninvasive and those that are more invasive. At present the noninvasive tests are generally offered to all pregnant women, while the more invasive tests are generally recommended only if some risk factors exist. The noninvasive tests include ultrasound imaging and maternal serum tests. Serum tests include one for alphafetoprotein (AFP) or one for alphafetoprotein, estriol, and human chorionic gonadotropin (triple screen). These tests serve as screens for structural fetal malformations and for neural tube closure defects. The triple screen also can detect some cases of Down syndrome, although there is a significant false positive and false negative rate.

Amniocentesis

More invasive tests include amniocentesis, chorionic villus sampling, percutaneous umbilical blood sampling, and, upon rare occasion, pre-implantation testing of either a polar body or a dissected embryonic cell. Amniocentesis is a procedure in which a long, thin needle is inserted through the abdomen and uterus into the amniotic sac, enabling the removal of a small amount of the amniotic fluid bathing the fetus. This procedure is generally performed under ultrasound guidance during the 15th to 17th week of pregnancy, and although it is generally regarded as safe, complications can occur, ranging from cramping to infection or loss of the fetus. The amniotic fluid obtained can be used in each of three ways: (1) living fetal cells recovered from this fluid can be induced to grow and can be analyzed to assess chromosome number, composition, or structure; (2) cells recovered from the fluid can be used for molecular studies; and (3) the amniotic fluid itself can be analyzed biochemically to determine the relative abundance of a variety of compounds associated with normal or abnormal fetal metabolism and development. Amniocentesis is typically offered to pregnant women over the age of 35 due to the significantly increased rate of chromosome disorders observed in the children of older mothers. A clear advantage of amniocentesis is the wealth of material obtained and the relative safety of the procedure. The disadvantage is timing: results may not be received until the pregnancy is already into the 19th week or beyond, at which point the possibility of termination may be much more physically and emotionally wrenching than if considered earlier.

Chorionic villus sampling (CVS) is a procedure in which either a needle is inserted through the abdomen or a thin tube is inserted into the vagina and cervix to obtain a small sample of placental tissue called chorionic villi. CVS has the advantage of being performed earlier in the pregnancy (generally 10–11 weeks), although the risk of complications is greater than that for amniocentesis. Risks associated with CVS include fetal loss and fetal limb reduction if the procedure is performed earlier than 10 weeks gestation. Another disadvantage of CVS reflects the tissue sampled: chorionic villi are not part of the embryo, and as such a sample may not accurately represent the embryonic genetic constitution. In contrast, amniotic cells are embryonic in origin, having been sloughed off into the fluid. Therefore, abnormalities, often chromosomal, may be seen in the chorionic villi but not in the fetus, or vice versa.

Both percutaneous umbilical blood sampling (PUBs) and pre-implantation testing are rare, relatively high-risk, and performed only in very unusual cases. Pre-implantation testing of embryos derived by in vitro fertilization is a particularly new technique and is currently used only in cases of couples at high risk for having a fetus affected with a given familial genetic disorder who find all other alternatives unacceptable. Pre-implantation testing involves obtaining eggs and sperm from the couple, combining them in the laboratory, and allowing the resultant embryos to grow until they reach the early blastocyst stage of development, at which point a single cell is removed from the rest and harvested for fluorescent in situ hybridization (FISH) or molecular analysis. The problem with this procedure is that one cell is scant material for diagnosis, so that a large array of tests cannot be performed. Similarly, if the test fails for any technical reason, it cannot be repeated. Finally, embryos determined to be normal and therefore selected for implantation into the mother are subject to other complications normally associated with in vitro fertilization—namely, that only a small fraction of the implanted embryos make it to term and that multiple, and therefore high-risk, pregnancies are common. Nonetheless, many at-risk couples find these complications easier to accept than an elective termination of the pregnancy.

Finally, it should be noted that researchers have identified fetal cells in the maternal circulation and that procedures are currently under development to enable their isolation and analysis, thereby providing a noninvasive alternative for molecular prenatal testing. Although these techniques are currently experimental and are not yet available for clinical application, they may well become the methods of choice in the future.

Genetic testing. In the case of genetic disease, options often exist for presymptomatic diagnosis—that is, diagnosis of individuals at risk for developing a given disorder, even though at the time of diagnosis they may be clinically healthy. Options may even exist for carrier testing, studies that determine whether an individual is at increased risk of having a child with a given disorder, even though he or she personally may never display symptoms. Accurate predictive information can enable early intervention, often preventing the clinical onset of symptoms rather than waiting for and then responding to them, in which case irreversible damage may have already occurred. In the case of carrier testing, accurate information can enable prospective parents to make more informed family planning decisions. Unfortunately, there can also be negative aspects of early detection, including such issues as privacy, individual responses to potentially negative information, discrimination in the workplace, or discrimination in access to or cost of health or life insurance. While some governments have outlawed the use of presymptomatic genetic testing information by insurance companies and employers, others have embraced it as a way to bring spiraling health care costs under control. Some communities have even considered instituting premarital carrier testing for common disorders in the populace.

Genetic testing procedures can be divided into two different groups: (1) testing of individuals considered at risk from phenotype or family history, and (2) screening of entire populations, regardless of phenotype or personal family history, for evidence of genetic disorders common in that population. Both forms are currently pursued in many societies. Indeed, with the explosion of information about the human genome and the increasing identification of potential "risk genes" for common disorders, such as cancer, heart disease, or diabetes, the role of predictive genetic screening in general medical practice is likely to increase.

At present, adults are generally tested for evidence of genetic disease only if personal or family history suggests they are at increased risk for a given disorder. A typical example would be a young man whose father, paternal aunt, and older brother have all been diagnosed with early onset colon cancer. Although this person may appear perfectly healthy, he is at significantly increased risk to carry mutations associated with familial colon cancer, and accurate genetic testing could enable heightened surveillance (*e.g.,* frequent colonoscopies) that might ultimately save his life.

Carrier testing for adults in most developed nations is generally offered only if family history or ethnic origins suggest an increased risk of having a particular disease. A typical example would be to offer carrier testing for cystic fibrosis to a couple including one member who has a sibling with the disorder. Another would be to offer carrier testing for Tay-Sachs disease to couples of Ashkenazic Jewish origin, a population known to carry an increased frequency of Tay-Sachs mutations. The same would be true for couples of African or Mediterranean descent with regard to sickle cell anemia or thalassemia, respectively. Typically, in each of these cases a genetic counselor would be involved to help the individuals or couples understand their options and make informed decisions.

Newborn screening

Screening of large unphenotyped populations for evidence of genetic disease is currently pursued in most industrialized nations only in the newborn population, although future developments in the identification of risk genes for common adult-onset disorders may change this policy. So-called mandated newborn screening was initiated in many societies in the latter quarter of the 20th century in an effort to prevent the drastic and often irreversible damage associated with a small number of relatively common genetic disorders whose sequelae can be either prevented or significantly relieved by early detection and intervention. The general practice is to collect a small sample of blood from each newborn, generally by pricking the infant's heel and collecting drops of blood on special filter paper, which is then analyzed. Perhaps the best-known disorder screened in this manner is phenylketonuria (PKU), an autosomal recessive inborn error of metabolism discussed in the section *Autosomal recessive inheritance.* With early diagnosis and dietary intervention that is maintained throughout life, children with PKU can escape mental retardation and grow into healthy adults who lead full and productive lives. Although many of the genetic disorders currently tested by mandated newborn screening are metabolic in nature, this trend is beginning to change. For example, in some communities newborns are screened for profound congenital hearing loss, which is now known to be frequently genetic in origin and for which effective intervention is now available (*e.g.,* through cochlear implants).

Genetic tests themselves can take many forms, and the choice of tests depends on a number of factors. For example, screening for evidence of sickle cell anemia, a hemoglobin disorder, is generally pursued at least initially by tests involving the hemoglobin proteins themselves, rather than DNA, because the relevant gene product (blood) is readily accessible and because the protein test is currently cheaper to perform than the DNA test. In contrast, screening for cystic fibrosis, a disorder that predominantly affects the lungs and pancreas, is generally pursued in the at-risk newborn at the level of DNA because there is no cheap and accurate alternative. Older persons suspected of having cystic fibrosis, however, can also be diagnosed with a "sweat test" that measures sweat electrolytes.

Tests involving analysis of DNA are particularly powerful because they can be performed using very tiny samples; also, the DNA tested can originate from almost any tissue type, regardless of whether the gene of interest happens to be expressed in that tissue. Current technologies applied for mutation detection include traditional karyotyping and Southern blotting, as well as a multitude of new tests, including FISH with specific probes or the polymerase chain reaction (PCR), which refers to an enzymatic process by which specific regions of the genome can be amplified for molecular study. Which tests are applied depends on whether the genetic abnormalities are likely to be chromosomal (in which case karyotyping or FISH is appropriate), large deletions or other rearrangements (best tested for by Southern blotting or PCR), or point mutations (best confirmed by PCR followed by oligonucleotide hybridization or restriction enzyme digestion). If a large number of different point mutations are sought, as is often the case, the most appropriate technology may be microarray hybridization analysis, which can test for tens to hundreds of thousands of different point mutations in the same sample simultaneously.

Options for treatment. Options for the treatment of genetic disease are both many and expanding. Although a significant number of genetic diseases still have no effective treatment, for many the treatments are quite good. Current approaches include dietary management, such as the restriction of phenylalanine in PKU; protein or enzyme replacement, such as that used in Gaucher's syndrome, hemophilia, and diabetes; and tissue replacement, such as blood transfusions or bone marrow transplantation in sickle cell anemia and thalassemia. Other treatments are strictly symptomatic, such as the use of splints in Ehlers-Danlos syndrome, administration of antibiotics in early cystic fibrosis, or female hormone replacement in Turner's syndrome. Many options involve surveillance and surgery, such as regular checks of aortic root diameter followed by surgery to prevent aortic dissection in Marfan's syndrome, or regular colonoscopy in persons at risk for familial colon cancer, followed by surgical removal of the colon at the first signs of disease.

Some genetic diseases may also be amenable to treatment by gene therapy, the introduction of normal genetic sequences to replace or augment the inherited gene whose mutation underlies the disease. Although some successes have been reported with gene therapy trials in humans—for example, with patients who have severe combined immunodeficiency (SCID) or hemophilia—significant technical challenges remain.

ETHICAL ISSUES

Our genetic constitution contributes to making us not only what we are—tall or short, male or female, healthy or sick—but also who we are—how we think and feel. Furthermore, although we generally like to think of our genomes as being uniquely ours, in fact we share significant aspects of them with our families, and information about our own genes is also information about our loved ones. Perhaps most important, in the biological sense, the genes we pass on to our children represent the closest we will ever come to immortality. For these reasons and others, human genetics is a topic fraught with ethical dilemma, with enormous power for good but also frightening possibilities for misuse. The challenge and responsibility are to harness available information and technologies to improve life and health for all people, without compromising privacy, autonomy, or diversity. Of vital importance in achieving these goals is an educated society that is aware of the advantages of new technologies yet is also concerned about their potential dangers.

(Ar.R./J.L.F.-K./I.F.)

BIBLIOGRAPHY

Classical genetics. THEODOSIUS DOBZHANSKY, *Heredity and the Nature of Men* (1964); THEODOSIUS DOBZHANSKY *et al., Evolution* (1977); and I. MICHAEL LERNER and WILLIAM J. LIBBY, *Heredity, Evolution, and Society,* 2nd ed. (1976), are excellent discussions of classical genetics and its social and cultural implications. CURT STERN, *Genetic Mosaics, and Other Essays* (1968), is a group of historical essays by a leading authority who discusses the development of knowledge on hermaphrodites and the relation of general to human genetics. JAMES A. PETERS (ed.), *Classic Papers in Genetics* (1959), is a collection of papers extending from 1865 (Mendel) to 1966 (Benzer) that form the cornerstone of classical Mendelian genetics. ARCHIBALD E. GARROD, *Inborn Errors of Metabolism* (1909, reprinted with a supplement by HARRY HARRIS, 1963), is a classic work. ROBERT OLBY, *Origins of Mendelism,* 2nd ed. (1985), provides a description of Mendel's experiments and discoveries placed in the context of hereditary thought before and during Mendel's era. A.H. STURTEVANT, *A History of Genetics* (1965, reissued 2001), is a review of the critical developments in the evolution of our understanding of heredity. JAMES D. WATSON, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA* (1968, reissued 2001), is written by one of the discoverers.

Genetic texts. ANTHONY J.F. GRIFFITHS *et al., Modern Genetic Analysis* (1999); ARTHUR P. MANGE and ELAINE JOHANSEN MANGE, *Genetics: Human Aspects,* 2nd ed. (1990); W.S. KLUG and M.R. CUMMINGS, *Essentials of Genetics,* 3rd ed. (1999); BENJAMIN LEWIN, *Genes VII* (2000); MICHAEL R. CUMMINGS, *Human Heredity: Principles and Issues,* 5th ed. (1999); RICKI LEWIS, *Human Genetics: Concepts and Applications,* 3rd ed. (1999); and DANIEL L. HARTL *Genetics: Analysis of Genes and Genomes,* 5th ed. (2001), are comprehensive tertiary-level genetics textbooks. VICTOR A. MCKUSICK, *Mendelian Inheritance in Man,* 12th ed., 3 vol. (1998), is a compendium of all the known inherited phenotypes with descriptions and references. GEORGE P. RÉDEI, *Genetics Manual: Current Theory, Concepts, Terms* (1998), provides a good up-to-date overview treatment for students.

JAMES D. WATSON *et al., Molecular Biology of the Gene,* 4th ed., 2 vol. (1987); and DAVID A. MICKLOS and GREG A. FREYER, *DNA Science: A First Course in Recombinant DNA Technology* (1990), provide an introduction to molecular biology; R.W. OLD and S.B. PRIMROSE, *Principles of Gene Manipulation: An Introduction to Genetic Engineering,* 5th ed. (1994), is one of the best texts dealing exclusively with genetic engineering. JAMES D. WATSON *et al., Recombinant DNA,* 2nd ed. (1992), provides a relatively short but extremely useful treatment of the methods and applications of recombinant DNA technology and other current methods in molecular genetics. PETER SUDBERY, *Human Molecular Genetics* (1998), describes some of the current technological approaches used in human genetics. HAROLD VARMUS and ROBERT A. WEINBERG, *Genes and the Biology of Cancer* (1998); and HARVEY LODISH *et al., Molecular Cell Biology,* 4th ed. (2000), are well-written books aimed at the general public. (Ar.R./A.J.F.G.)

Geneva

One of Europe's most cosmopolitan cities, Geneva (French Genève, German Genf, Italian Ginevra) has served as a model for republican government and owes its preeminence to the triumph of human, rather than geographic, factors. It developed its unique character from the 16th century, when, as the centre of the Calvinist Reformation, it became the "Protestant Rome." Situated in the far southwestern corner of Switzerland that juts into France, the canton of Geneva has a total area of 109 square miles (282 square kilometres), of which seven square miles constitute the city proper. Territorial isolation has been a basic feature of this region, which did not establish its definitive frontiers until 1815. Cut off politically and culturally after the Reformation from its natural geographic surroundings in Roman Catholic France and Savoy, Geneva was forced to establish an attenuated but powerful network of intellectual and economic relationships with the rest of Europe and with nations overseas.

A city-state transformed after many vicissitudes into a democratic Swiss canton, Geneva has functioned primarily as a centre of commerce, in contact with both Germanic and Mediterranean countries. Contemporary Geneva is, above all, a service metropolis, retaining its financial importance and housing the headquarters of many public and private international organizations.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

Site. Geneva is located at the southwestern end of Lake Geneva (Lac Léman) at its junction with the Rhône River. The city lies at an elevation of 1,230 feet (375 metres) in the centre of a natural basin encircled by mountains. This excellent site, besides commanding the important Swiss corridor between the Alps and the Jura Mountains, is also the focus of Alpine passes leading into Italy and, along the Saône–Rhône axis, of routes to the Mediterranean.

Climate. The local climate is tempered by the presence of the lake, while the Jura create a screen that diminishes rainfall. Average temperatures in Geneva are about 32° F (0° C) in January and about 64° F (18° C) in July. Geneva is thus neither disagreeably hot in summer nor cold in winter, but it must sometimes endure the harsh north wind known as the bise. Annual precipitation averages about 37 inches (930 millimetres).

Layout. Bisected by the lower lake basin and the river, Geneva exhibits the classic pattern of old European cities, with neighbourhoods lying in belts around the original nucleus. The Haute-ville, or upper city, centred on the city's original hill site at the Plateau des Tranchées and dominated by the Cathedral of St. Peter, is the historic heart of Geneva. The typical medieval and Renaissance houses are crowded together along narrow streets. This neighbourhood has undergone relative depopulation as housing has given way to government buildings and art, antiques, and interior furnishings businesses.

Residential areas

At the foot of the hill an area reclaimed from the lake and the Rhône forms a low-lying shopping district. On the site of the old fortifications—mostly to the south of the Rhône—lie suburbs dating from the 19th century. Beyond is an irregular belt of working-class residential areas, near the railway stations and industrial zones.

International agencies such as the Red Cross and the World Health Organization are found on the old patrician properties north of the Rhône. In this section, too, is the Palais des Nations, now the European home of the United Nations. At the lake's edge the Jet d'Eau—reputedly the world's tallest fountain, with a jet of water rising 476 feet (145 metres)—provides a familiar symbol of the city.

THE PEOPLE

It was not until after 1945 that the city's population began to register rapid growth, with the influx of other Swiss citizens and foreigners attracted by Geneva's international institutions and financial, chemical, and construction industries. By the late 1980s the population was approximately one-third foreign, one-third Swiss from other cantons, and only one-third native Genevese. Immigration to Geneva has consisted not only of the traditional contingents from Italy, France, and the Iberian Peninsula but also of a rising number from the Americas, Asia, and Africa. Although the large foreign presence is one of the constants of the city's demography, French remains the first language of Geneva.

Among the native population and in the professional classes, Protestants are in the majority, but within the population as a whole, Geneva is no longer the "Protestant Rome." Roman Catholics, in fact, make up slightly more than half the population.

THE ECONOMY

Industry. Manufacturing is handicapped by lack of space and raw materials, but Geneva, as one of the oldest banking centres in Europe, has profited from an early start in capital accumulation. It benefits from a skilled labour force and managers who are international in outlook. Certain older activities, such as cotton textile manufacture, have disappeared, but watchmaking has a continuing tradition of precision and quality. Industrial production is diversified and is, above all, designed for export. The largest industry is the manufacture of instruments and precision machinery. Principal specialties are equipment for hydraulic plants (turbines and alternators), electrical equipment, machine tools, and measuring devices.

The chemical industry is the second largest in Switzerland, after that of Basel. It supplies luxury items—such as fragrances and bases for perfume—as well as medicines. The food-processing industry is important, and Geneva also manufactures almost half of all Swiss-made cigarettes. Agriculture supplies such commodities as wheat, rapeseed, dairy products, and wine. Only about 1 percent of the canton's people are employed in farming.

Commerce and finance. Service industries employ more

Geneva in its lake setting, with the International Labour Office and the Parc Villa Barton in the foreground.

By courtesy of the Swiss National Tourist Office

than two-thirds of the population. Wholesale and retail trade, banking, tourism, insurance, and the stock exchange are among the principal employers. Although nationally Geneva is second to Zürich in total volume of financial transactions, it has retained a position of worldwide significance. Geneva is one of the world's leading sanctuaries for capital, and it has been estimated that its banks hold more than half the total amount of foreign capital in Switzerland.

Financial role

Transportation. In the area of transport, success came late. It was said that Geneva lost to Lausanne the battle to become a leading railroad centre in the 19th century, but since World War II the city has acquired a large international airport at Cointrin. Multilane expressways have linked Geneva with Lausanne and with the rest of the Swiss highway system since 1964 and with the French system since 1970. In addition, the city contributed labour and financing for the construction of the highway tunnel beneath Mont Blanc and the Route-Blanche (White Way) to Italy. Since 1984 Geneva has enjoyed a high-speed railway system, the *trains à grande vitesse* (TGV), providing a three-hour connection with Paris. Local transportation is provided by an extensive bus, trolley, and streetcar system.

ADMINISTRATION

The canton of Geneva, which still calls itself La Republique du Genève, is governed by the Constitution of 1848 (as amended). Cantonal government is exercised by an executive power, the Council of State, consisting of seven members who are elected for four-year terms, and by a legislature, the Great Council, composed of 100 deputies who are also elected for four-year terms by proportional ballot.

The canton is divided into communes, each of which has its own assembly, administrative council, and mayor. Citizens have the rights of legislative initiative and referendum at both the communal and cantonal levels. To represent it in the federal government, the canton elects two deputies to the Council of States and a varying number of representatives to the National Council.

CULTURAL LIFE

Geneva has an ancient cultural tradition. A scholarly elite long cultivated theology, philosophy, literature, and, especially since the 17th century, the natural and applied sciences. Numerous scientific organizations are based in Geneva, including the European Organization for Nuclear Research (CERN), a leader in subnuclear physics research, and the World Meteorological Organization (WMO). The Geneva City Conservatory and Botanical Gardens is a major botanical research centre. In 1872 the Academy, in existence since the 16th century, became a university, and it has acquired an outstanding reputation. Other aspects of Geneva's active cultural life revolve around its museums, the Grand Théâtre (the city's opera house), and the proceedings of international meetings held there. The music conservatory and international performance competitions attract large numbers of musicians, and the Orchestre de la Suisse Romande is renowned worldwide. There are a number of distinguished small publishing houses in Geneva, and the city contributes substantially to the French-language services of the Swiss television and radio system. The *Journal de Geneve,* one of the world's premier newspapers, merged with *Le Nouveau Quotidien* to become *Le Temps* in 1998.

The lake provides many recreational opportunities for swimming, sailing, and fishing. Winter sports such as skiing and skating are popular, and rock climbing and mountaineering are pursued for both science and sport.

History

FOUNDATION AND MEDIEVAL GROWTH

The original site of the city was an easily defended hill dominating the outlet of the lake. Human occupation began in the Paleolithic Period and further developed in the Neolithic, which was marked by the growth of a vast lake-dwelling community with habitations built on piles. The original name of Genava (or Geneva) undoubtedly dates back to the pre-Celtic Ligurian peoples. In about 500 BC Geneva was a fortified settlement of the Allobrogian Celts, and as early as 58 BC it served as a departure point in the campaign of the Helvetians and the Romans for Gaul. By AD 379 Geneva was the seat of a bishop and was within the Roman Empire, but when it had been Christianized and when it became a Roman city are uncertain. After the Germanic invasions Geneva became

Capital of Burgundy

part of the Burgundian kingdom and served as its first capital from 443 to 534.

For a time Geneva belonged to Lotharingia and then again to Burgundy (888–1032). During the early feudal period the city formed the hub of the lands belonging to the Genevese counts. With the final extinction of their line in 1401, the bishop, who was a direct vassal of the Holy Roman emperor and invested with temporal power, vied for control with the neighbouring counts of Savoy.

THE 15TH TO 18TH CENTURY

In the 15th century the counts of Savoy rose to the status of dukes and made strenuous efforts to assert their sovereignty in Geneva at the expense of the bishops, who made correspondingly generous offers to the burghers to win their support against the dukes. But the burghers were slow to forsake the dukes, from whom they secured a contract recognizing their General Council—the public assembly to which every citizen belonged—as the central legislative body of the city.

Geneva and Savoy. The dukes of Savoy were ambitious and successful rulers who in time assumed a kingly title. They continued to assert their claims to Geneva, even when it lost to Lyon its preeminence as a centre of international trade fairs, with the result that its prosperity and population declined. The dukes used cunning as well as force to uphold their sovereignty, and from 1449 until 1522 they had members of their own family enthroned as bishop of Geneva.

Independence

The last ruling bishop, Pierre de La Baume, fled from Geneva in July 1533, and a year later the burghers declared the see vacant. Thus they rid themselves at once of their bishop and their allegiance to Savoy, and proclaimed themselves a state. When the Savoyards threatened invasion a year later, the Bernese offered to incorporate Geneva under their government. Having no wish to exchange the domination of Savoy for that of Bern, the Genevans refused. Because they desperately needed Bernese troops, however, they could not safely object to a rapprochement with Protestant Bern in the matter of religion; so in 1536 they declared themselves Protestant, a move that also served to justify the permanent exclusion of the bishop. As a result, they alienated the Roman Catholic Swiss cantons, so that Geneva's adhesion to the confederation was vetoed for generations to come.

John Calvin. Protestantism did not appeal immediately to everyone in Geneva. Some felt closer to French-speaking, Roman Catholic Fribourg than to relatively patrician, German-speaking Bern; and for many the theology of Martin Luther and Huldrych Zwingli was altogether foreign. This situation was resolved by John Calvin, a French theologian and practical visionary who transformed Geneva into a modern city-state and reconciled its people to the Reformed religion. Adapting traditional institutions to serve new purposes, Calvin was remarkably successful in presiding over Geneva's formative years as an autonomous state. He owed his success in part to the continuing presence of the Protestant Bernese troops. He was thus able to reorganize Geneva without hostile intervention by the Roman Catholic Savoyards, whose forces at other times stood on the frontiers of the city.

Calvin was also fortunate in that the persecution of Protestants in France brought into Geneva refugees sympathetic to his purposes. This enabled him to replenish with immigrants a citizen roll diminished by his own harsh policy of expelling all those who resisted conversion to the Reformed religion. The immigrants brought new trades, industries, and wealth, and Geneva became an industrial, financial, and commercial metropolis. Calvin's academies and seminaries attracted scholars from all over Europe.

A few such visitors found that they had only exchanged one form of persecution for another. The Spanish-born physician and theological writer Michael Servetus and Jacques Gruet, an apostate Protestant, were put to death for heresy. As Geneva grew and prospered, however, religious fanaticism died down.

The Savoyards made a final abortive attempt to recapture Geneva with a surprise attack led by the Duke on the night of Dec. 11–12, 1602, but they were driven out in a brief skirmish. This event, known as the Escalade, is still commemorated annually in Geneva.

Class conflicts. Between the mid-16th and early 18th centuries, the powers of the aristocratic Council of Twenty-five were systematically enlarged at the expense of the General Council, which eventually was summoned only to rubber-stamp the decisions of the magistrates.

Social changes added a further dimension to these developments. Among the French and Italian Protestants who found refuge in Geneva were several from noble families who brought with them not only their wealth but also their assumed right to lead and rule. These families grew to monopolize the Council of Twenty-five and to set up what was in fact the rule of a hereditary nobility, but one veiled by the ceremonies, styles, and language of republicanism.

Social changes

Social change of another kind was taking place as well. The number of residents of Geneva who were able to qualify as citizens became proportionately smaller as the population grew from about 13,000 to 25,000. In the 16th century the great majority of male residents were citizens; by 1700 the citizens constituted a minority—only about 1,500 of Geneva's 5,000 adult males. The other inhabitants were not only excluded from many civil rights and privileges but also were denied access to all the most lucrative trades and professions.

For reasons such as these, discontented factions multiplied behind the tranquil facade of Genevan life. There were citizens who opposed the domination of the patrician families, and there were unenfranchised residents who opposed the monopoly of rights and privileges by the citizens. Opposition to the ruling clique developed among the citizens at the end of the 17th century, asserting the rights of the General Council against the usurpations of the Council of Twenty-five.

Despite these currents of political opposition, Geneva in the 18th century was at the zenith of its prosperity. Material wealth stimulated a burst of culture and artistic creativity. As the birthplace of Rousseau and the sanctuary of Voltaire, Geneva attracted the elite of the Enlightenment and helped to foster the development of the new political science, derived from natural law.

In 1798, with the aid of local Jacobins, Geneva was annexed to France. The city was reduced to a subservient role and submitted, in 1802, to the protection of Napoleon Bonaparte. The Emperor distrusted Geneva, "that city where they know English too well" (it was indeed harbouring a secret liberal and Anglophile opposition), and the French period became an era of stagnation and recession.

THE 19TH AND 20TH CENTURIES

Swiss Geneva. As early as 1813 Geneva threw in its lot with France's enemies and was thus able to claim indemnities upon the fall of the empire. The aristocratic republic was restored and undertook negotiations to join the Swiss Confederation. On Sept. 12, 1814, the Genevan republic was admitted to the ranks of the Swiss cantons. Through the cession of 12 Savoyard communes by the Second Treaty of Paris (Nov. 20, 1815), it rounded out its territories into a single block.

Geneva's aristocrats were again in power, and gradually the bourgeoisie and the common people began once more to challenge openly the patrician regime. On Oct. 7, 1846, the working-class suburb of Saint-Gervais revolted, and the conservative government was overthrown. Opposition by the Swiss Diet to the Sonderbund (a league of seven Roman Catholic cantons) and the 1847 civil war between federal forces and the rebellious cantons permitted the radicals, led by James Fazy, to take the offensive. The radicals, who drew up the new Constitution of 1848, were thereafter masters of Geneva, and Fazy dominated the political scene until 1861. In many ways the founder of modern Geneva, he opened the canton to railway lines, created the Bank of Geneva, and, above all, made widespread urban expansion possible by demolishing the city's outer fortifications.

In 1860 the Savoyards voted to accept the sovereignty of France, and a free zone was created for Geneva by agreement with the French. The city regained, and until 1914 held, its role as a regional economic capital. It also continued to assert its international influence. The Red Cross was founded in Geneva in 1864; the Geneva conventions for the protection of prisoners of war were signed there; and the League of Nations was installed in the city in 1919.

The city since 1945. The history of Geneva since World War II has been marked by steady economic growth, halted only temporarily by the oil crisis of the early 1970s. The population of the canton increased from 187,000 in 1945 to more than 350,000 by the mid-1980s, and revenues rose from 660,000,000 Swiss francs to more than 9,000,000,000 during the same period. This prosperity was experienced almost entirely in the commercial and financial sectors; industry declined radically, affording employment to only 20 percent of the work force in 1980, as opposed to more than 36 percent in 1950. Building alone among Geneva's industries flourished after the war, as offices, houses, and shops—indeed whole new suburbs—had to be provided for the ever-increasing population.

In keeping with its cosmopolitan traditions, Geneva attracted international bodies seeking a location for their headquarters. The United Nations took over the old League of Nations buildings; the International Labour Organisation, the World Council of Churches, and other institutions resumed their operations in Geneva; and the city became a favoured neutral meeting place for diplomatic initiatives.

Postwar politics

In 1960 Geneva was one of the first Swiss cantons to extend the vote to women, but participation in elections and referendums remained unusually low. Genevese political parties were generally to the left of their counterparts in the Confederation, but they continued to maintain consensus politics and coalition government; this occurred despite the challenge of the communists, legalized as the Workers' Party (Parti du Travail) in 1944, and the right-wing nationalists, or Vigilantes, who had some success in the elections of 1965. In the federal government at Bern, representatives of Geneva failed to attain much prominence, and political life in Geneva tended to be centred more on the canton than on the nation. (P.G./M.C.)

For most of the post-World War II era, Geneva experienced continuous economic growth as international organizations and companies built headquarters in the city. However, during the late 1980s and early '90s the city began to stagnate as some international organizations left and the real-estate bubble, which had fueled a dramatic increase in property prices, burst. Throughout much of the 1990s the city's economy lagged behind the rest of Switzerland, and the unemployment rate, which hitherto had been negligible, was among the highest in the country. By the end of the 1990s, the economy had begun to recover.

Despite increasing competition from other cities, Geneva maintained its reputation as an international city throughout the last decades of the 20th century. In 1979 Geneva became the permanent headquarters for the international Disarmament Conference, involving more than 60 countries. A nuclear test ban treaty and an agreement to prohibit the production of antipersonnel mines were among the conference's breakthroughs. Geneva was also the site of the historic initial summit between U.S. President Ronald Reagan and Soviet leader Mikhail Gorbachev in 1985. Although the meeting did not produce any firm commitments, it was the first time the leaders had discussed nuclear arms reductions and paved the way for later agreements and the eventual end of the Cold War. Under the auspices of the United Nations Children's Fund, the International Convention on the Rights of Children was negotiated in Geneva in 1994. In 1995 the World Trade Organization was established with Geneva as its headquarters. The city is also the headquarters for the European Organization for Nuclear Research (CERN), which is commonly credited with developing the World Wide Web. By the beginning of the 21st century, international organizations based in Geneva had been selected as Nobel Prize winners more than 40 times. With Switzerland's neutral foreign policy, Geneva was expected to continue its central international role well into the 21st century. (Ed.)

BIBLIOGRAPHY. Good introductions to Geneva are provided by the collective volume, BENJAMIN LAEDERER (ed.), *Geneva: Crossroad of the Nations* (1964; originally published in French, 1963); and PAUL GUICHONNET, *Histoire de Genève* (1974). Urban evolution is the subject of LOUIS BLONDEL, *Le Développement urbain de Genève à travers les siècles* (1946); and ANDRÉ CORBOZ, *Invention de Carouge, 1772–1792* (1968). The historical literature is rich, although little is available in English translation. PAUL-F. GEISENDORF, *Bibliographie raisonnée de l'histoire de Genève des origines à 1798* (1966), is a well-researched bibliography. See also LOUIS BINZ, *Genève et les Suisses: du Moyen Age à la Restauration* (1964), a detailed survey, and *Brève histoire de Genève* (1981); RENÉ GUERDAN, *Histoire de Genève* (1981), especially useful for the modern period; PAUL-E. MARTIN (ed.), *Histoire de Genève,* 2 vol. (1951–56); and FRANÇOIS RUCHON, *Histoire politique de la République de Genève . . . ,* 2 vol. (1953). For economic history, see ANTONY BABEL, *Histoire économique de Genève, des origines au début du XVI*^{e} *siècle,* 2 vol. (1963); JEAN-FRANÇOIS BERGIER, *Genève et l'économie européenne de la Renaissance* (1963); and ANNE MARIE PIUZ, *Affaires et politique: recherches sur le commerce de Genève au XVII*^{e} *siècle* (1964). See also WALDEMAR DEONNA, *Les Arts à Genève des origines à la fin du XVIII*^{e} *siècle* (1942); and ALFRED BERCHTOLD, *La Suisse romande au cap du XX*^{e} *siècle: portrait littéraire et moral* (1963). (P.G./M.C.)

Genghis Khan

The Mongol Temüjin, known to history as Genghis Khan (Chinggis, or Jenghiz, Khan), was a warrior and ruler of genius who, starting from obscure and insignificant beginnings, brought all the nomadic tribes of Mongolia under the rule of himself and his family in a rigidly disciplined military state. He then turned his attention toward the settled peoples beyond the borders of his nomadic realm and began the series of campaigns of plunder and conquest that eventually carried the Mongol armies as far as the Adriatic Sea in one direction and the Pacific coast of China in the other, leading to the establishment of the great Mongol Empire.

By courtesy of the National Palace Museum, Taipei, Taiwan, Republic of China

Genghis Khan, ink and colour on silk. In the National Palace Museum, Taipei.

Historical background. With the exception of the saga-like *Secret History of the Mongols* (1240?), only non-Mongol sources provide near-contemporary information about the life of Genghis Khan. Almost all writers, even those who were in the Mongol service, have dwelt on the enormous destruction wrought by the Mongol invasions. One Arab historian openly expressed his horror at the recollection of them. Beyond the reach of the Mongols and relying on second-hand information, the 13th-century chronicler Matthew Paris called them a "detestable nation of Satan that poured out like devils from Tartarus so that they are rightly called Tartars." He was making a play on words with the classical word Tartarus (Hell) and the ancient tribal name of Tatar borne by some of the nomads, but his account catches the terror that the Mongols evoked. As the founder of the Mongol nation, the organizer of the Mongol armies, and the genius behind their campaigns, Genghis Khan must share the reputation of his people, even though his generals were frequently operating on their own, far from direct supervision. Nevertheless, it would be mistaken to see the Mongol campaigns as haphazard incursions by bands of marauding savages. Nor is it true, as some have supposed, that these campaigns were somehow brought about by a progressive desiccation of Inner Asia that compelled the nomads to look for new pastures. Nor, again, were the Mongol invasions a unique event. Genghis Khan was neither the first nor the last nomadic conqueror to burst out of the steppe and terrorize the settled periphery of Eurasia. His campaigns were merely larger in scale, more successful, and more lasting in effect than those of other leaders. They impinged more violently upon those sedentary peoples who had the habit of recording events in writing, and they affected a greater part of the Eurasian continent and a variety of different societies.

Struggles between nomadic and sedentary societies

Two societies were in constant contact, two societies that were mutually hostile, if only because of their diametrically opposed ways of life, and yet these societies were interdependent. The nomads needed some of the staple products of the south and coveted its luxuries. These could be had by trade, by taxing transient caravans, or by armed raids. The settled peoples of China needed the products of the steppe to a lesser extent, but they could not ignore the presence of the nomadic barbarians and were forever preoccupied with resisting encroachment by one means or another. A strong dynasty, such as the 17th-century Manchu, could extend its military power directly over all Inner Asia. At other times the Chinese would have to play off one set of barbarians against another, transferring their support and juggling their alliances so as to prevent any one tribe from becoming too strong.

The cycle of dynastic strength and weakness in China was accompanied by another cycle, that of unity and fragmentation amongst the peoples of the steppe. At the peak of their power, a nomadic tribe under a determined leader could subjugate the other tribes to its will and, if the situation in China was one of weakness, might extend its power well beyond the steppe. In the end this extension of nomadic power over the incompatible, sedentary culture of the south brought its own nemesis. The nomads lost their traditional basis of superiority—that lightning mobility that required little in the way of supply and fodder—and were swallowed up by the Chinese they had conquered. The cycle would then be resumed; a powerful China would reemerge, and disarray and petty squabbling among ephemeral chieftains would be the new pattern of life among the nomads. The history of the Mongol conquests illustrates this analysis perfectly, and it is against this background of political contrasts and tensions that the life of Genghis Khan must be evaluated. His campaigns were not an inexplicable natural or even God-given catastrophe but the outcome of a set of circumstances manipulated by a soldier of ambition, determination, and genius. He found his tribal world ready for unification, at a time when China and other settled states were, for one reason or another, simultaneously in decline, and he exploited the situation.

Early struggles. Various dates are given for the birth of Temüjin (or Temuchin), as Genghis Khan was named—after a leader who was defeated by his father, Yesügei, when Temüjin was born. The chronology of Temüjin's early life is uncertain. He may have been born in 1155, in 1162 (the date favoured today in Mongolia), or in 1167. According to legend, his birth was auspicious, because he came into the world holding a clot of blood in his hand. He is also said to have been of divine origin, his first ancestor having been a gray wolf, "born with a destiny from heaven on high." Yet his early years were anything but promising. When he was nine, Yesügei, a member of the royal Borjigin clan of the Mongols, was poisoned by a band of Tatars, another nomadic people, in continuance of an old feud.

Cast out by his clan

With Yesügei dead, the remainder of the clan, led by the rival Taychiut family, abandoned his widow, Höelün, and her children, considering them too weak to exercise leadership and seizing the opportunity to usurp power. For a time the small family led a life of extreme poverty, eating roots and fish instead of the normal nomad diet of mutton and mare's milk. Two anecdotes illustrate both Temüjin's straitened circumstances and, more significantly, the power he already had of attracting supporters through sheer force of personality. Once he was captured by the Taychiut, who, rather than killing him, kept him around their camps, wearing a wooden collar. One night, when they were feasting, Temüjin, noticing that he was being ineptly guarded, knocked down the sentry with a blow from his wooden collar and fled. The Taychiut searched all night for him, and he was seen by one of their people, who, impressed by the fire in his eyes, did not denounce him but helped him escape at the risk of his own life. On another occasion horse thieves came and stole eight of the nine horses that the small family owned. Temüjin pursued them. On the way he stopped to ask a young stranger, called Bo'orchu, if he had seen the horses. Bo'orchu immediately left the milking he was engaged in, gave Temüjin a fresh horse, and set out with him to help recover the lost beasts. He refused any reward but, recognizing Temüjin's authority, attached himself irrevocably to him as a *nökör,* or free companion, abandoning his own family.

Temüjin and his family apparently preserved a considerable fund of prestige as members of the royal Borjigin clan, in spite of their rejection by it. Among other things, he was able to claim the wife to whom Yesügei had betrothed him just before his death. But the Merkit people, a tribe living in north Mongolia, bore Temüjin a grudge, because Yesügei had stolen his own wife, Höelün, from one of their men, and in their turn they ravished Temüjin's wife Börte. Temüjin felt able to appeal to Toghril, Khan of the Kereit tribe, with whom Yesügei had had the relationship of *anda,* or sworn brother, and at that time the most powerful Mongol prince, for help in recovering Börte. He had had the foresight to rekindle this friendship by presenting Toghril with a sable skin, which he himself had received as a bridal gift. He seems to have had nothing else to offer; yet, in exchange; Toghril promised to reunite Temüjin's scattered people, and he is said to have redeemed his promise by furnishing 20,000 men and persuading Jamuka, a boyhood friend of Temüjin's, to supply an army as well. The contrast between Temüjin's destitution and the huge army furnished by his allies is hard to explain, and no authority other than the narrative of the *Secret History* is available.

Rise to power. With powerful allies and a force of his own, Temüjin routed the Merkit, with the help of a strategy by which Temüjin was regularly to scotch the seeds of future rebellion. He tried never to leave an enemy in his rear; years later, before attacking China, he would first make sure that no nomad leader survived to stab him in the back. Not long after the destruction of the Merkit, he treated the nobility of the Jürkin clan in the same way. These princes, supposedly his allies, had profited by his absence on a raid against the Tatars to plunder his property. Temüjin exterminated the clan nobility and took the common people as his own soldiery and servants. When his power had grown sufficiently for him to risk a final showdown with the formidable Tatars, he first defeated them in battle and then slaughtered all those taller than the height of a cart axle. Presumably the children could be expected to grow up ignorant of their past identity and to become loyal followers of the Mongols. When the alliance with Toghril of the Kereit at last broke down and Temüjin had to dispose of this obstacle to supreme power, he dispersed the Kereit people among the Mongols as servants and troops. This ruthlessness was not mere wanton cruelty. Temüjin intended to leave alive none of the old, rival aristocrats, who might prove a focus of resistance; to provide himself with a fighting force; and, above all, to crush the sense of clan loyalties that favoured fragmentation and to unite all the nomads in personal obedience to his family. And when, in 1206, he was accepted as emperor of all the steppe people, he was to distribute thousands of families to the custody of his own relatives and companions, replacing the existing pattern of tribes and clans by something closer to a feudal structure.

Destruction of the Merkit

At least from the time of the defeat of the Merkits, Temüjin was aiming at supremacy in the steppes for himself. The renewed friendship with Jamuka lasted only a year and a half. Then, one day while the two friends were on the march, Jamuka uttered an enigmatic remark about the choice of camping site, which provoked Temüjin's wife Börte to advise him that it was high time for the two friends to go their separate ways. What lies behind this episode is difficult to see. The story in the *Secret History* is too puzzling in its brevity and its allusive language to permit a reliable explanation. It has been suggested that Jamuka was trying to provoke a crisis in the leadership. Equally, it may be that the language is deliberately obscure to gloss over the fact that Temüjin was about to desert his comrade. In any event, Temüjin took Börte's advice. Many of Jamuka's own men also abandoned him, probably seeing in Temüjin the man they thought more likely to win in the end. The *Secret History* justifies their action in epic terms. One of the men tells Temüjin of a vision that had appeared to him and that could only be interpreted as meaning that Heaven and Earth had agreed that Temüjin should be lord of the empire. Looking at the situation in a more down-to-earth way, the interplay of the vacillating loyalties of the steppe may be discerned. The clansmen knew what was afoot, and some of them hastened to move over to Temüjin's side, realizing that a strong leader was in the offing and that it would be prudent to declare for him early on.

The break with Jamuka brought about a polarization within the Mongol world that was to be resolved only with the disappearance of one or the other of the rivals. Jamuka has no advocate in history. The *Secret History* has much to tell about him, not always unsympathetically, but it is essentially the chronicle of Temüjin's family; and Jamuka appears as the enemy, albeit sometimes a reluctant one. He is an enigma, a man of sufficient force of personality to lead a rival coalition of princes and to get himself elected *gur-khān,* or supreme Khan, by them. Yet he was an intriguer, a man to take the short view, ready to desert his friends, even turn on them, for the sake of a quick profit. But for Temüjin, it might have been within Jamuka's power to dominate the Mongols, but Temüjin was incomparably the greater man; and the rivalry broke Jamuka.

Gathering of the clans around Temüjin and Jamuka

Clan leaders began to group themselves around Temüjin and Jamuka, and, a few years before the turn of the century, some of them proposed to make Temüjin Khan of the Mongols. The terms in which they did so, promising him loyalty in war and the hunt, suggest that all they were looking for was a reliable general, certainly not the overlord he was to become. Indeed, later on, some of them were to desert him. Even at this time, Temüjin was only a minor chieftain, as is shown by the next important event narrated by the *Secret History,* a brawl at a feast, provoked by his nominal allies the Jürkin princes, whom he later massacred. The Chin emperor in north China, too, looked on him as of no great consequence. In one of the reversals of policy characteristic of their manipulation of the nomads, the Chin attacked their onetime allies the Tatars. Together with Toghril, Temüjin seized the opportunity of continuing the clan feud and took the Tatars in the rear. The Chin emperor rewarded Toghril with the Chinese title of *wang,* or prince, and gave Temüjin an even less exalted one. And, indeed, for the next few years the Chin had nothing to fear from Temüjin. He was fully occupied in building up his power in the steppe and posed no obvious threat to China.

Temüjin now set about systematically eliminating all rivals. Successive coalitions formed by Jamuka were defeated. The Tatars were exterminated. Toghril allowed himself to be manoeuvred by Jamuka's intrigues and by his own son's ambitions and suspicions into outright war against Temüjin, and he and his Kereit people were destroyed. Finally, in the west, the Naiman ruler, fearful of the rising power of the Mongols, tried to form yet another coalition, with the participation of Jamuka, but was utterly defeated and lost his kingdom. Jamuka, inconstant as ever, deserted the Naiman Khan at the last moment. These campaigns took place in the few years before 1206 and left Temüjin master of the steppes. In that year a great assembly was held by the River Onon, and Temüjin was proclaimed Genghis Khan: the title probably meant Universal Ruler.

Unification of the Mongol nation. The year 1206 was a turning point in the history of the Mongols and in world history: the moment when the Mongols were first ready to move out beyond the steppe. Mongolia itself took on a new shape. The petty tribal quarrels and raids were a thing of the past. Either the familiar tribe and clan names had fallen out of use or those bearing them were to be found, subsequently, scattered all over the Mongol world, testifying to the wreck of the traditional clan and tribe system. A unified Mongol nation came into existence as the personal creation of Genghis Khan and, through many vicissitudes (feudal disintegration, incipient retribalization, colonial occupation), has survived to the present day. Mongol ambitions looked beyond the steppe. Genghis Khan was ready to start on his great adventure of world conquest. The new nation was organized, above all, for war. Genghis Khan's troops were divided up on the decimal system, were rigidly disciplined, and were well equipped and supplied. The generals were his own sons or picked men, absolutely loyal to him.

Beginning of the conquest of Asia

Genghis Khan's military genius could adapt itself to rapidly changing circumstances. Initially his troops were exclusively cavalry, riding the hardy, grass-fed Mongol pony that needed no fodder. With such an army, other nomads could be defeated, but cities could not be taken. Yet before long the Mongols were able to undertake the siege of large cities, using mangonels, catapults, ladders, burning oil, and so forth and even diverting rivers. It was only gradually, through contact with men from the more settled states, that Genghis Khan came to realize that there were more sophisticated ways of enjoying power than simply raiding, destroying, and plundering. It was a minister of the Khan of the Naiman, the last important Mongol tribe to resist Genghis Khan, who taught him the uses of literacy and helped reduce the Mongol language to writing. The *Secret History* reports it was only after the war against the Muslim empire of Khwārezm, in the region of the Amu Darya (Oxus) and Syr-darya (Jaxartes), probably in late 1222, that Genghis Khan learned from Muslim advisers the "meaning and importance of towns." And it was another adviser, formerly in the service of the Chin emperor, who explained to him the uses of peasants and craftsmen as producers of taxable goods. He had intended to turn the cultivated fields of north China into grazing land for his horses.

Subjugation of China

The great conquests of the Mongols, which would transform them into a world power, were still to come. China was the main goal. Genghis Khan first secured his western flank by a tough campaign against the Tangut kingdom of Hsi Hsia, a northwestern border state of China, and then fell upon the Chin empire of north China in 1211. In 1214 he allowed himself to be bought off, temporarily, with a huge amount of booty, but in 1215 operations werc resumed, and Peking was taken. Subsequently, the more systematic subjugation of north China was in the hands of his general Muqali. Genghis Khan himself was compelled to turn aside from China and carry out the conquest of Khwārezm. This war was provoked by the governor of the city of Otrar, who massacred a caravan of Muslim merchants who were under Genghis Khan's protection. The Khwārezm-Shāh refused satisfaction. War with Khwārezm would doubtless have come sooner or later, but now it could not be deferred. It was in this war that the Mongols earned their reputation for savagery and terror. City after city was stormed, the inhabitants massacred or forced to serve as advance troops for the Mongols against their own people. Fields and gardens were laid waste and irrigation works destroyed as Genghis Khan pursued his implacable vengeance against the royal house of Khwārezm. He finally withdrew in 1223 and did not lead his armies into war again until the final campaign against Hsi Hsia in 1226–27. He died on August 18, 1227.

Assessment. As far as can be judged from the disparate sources, Genghis Khan's personality was a complex one. He had great physical strength, tenacity of purpose, and an unbreakable will. He was not obstinate and would listen to advice from others, including his wives and mother. He was flexible. He could deceive but was not petty. He had a sense of the value of loyalty, unlike Toghril or Jamuka. Enemies guilty of treachery toward their lords could expect short shrift from him, but he would exploit their treachery at the same time. He was religiously minded, carried along by his sense of a divine mission, and in moments of crisis he would reverently worship the Eternal Blue Heaven, the supreme deity of the Mongols. So much is true of his early life. The picture becomes less harmonious as he moves out of his familiar sphere and comes into contact with the strange, settled world beyond the steppe. At first he could not see beyond the immediate gains to be got from massacre and rapine and, at times, was consumed by a passion for revenge. Yet all his life he could attract the loyalties of men willing to serve him, both fellow nomads and civilized men from the settled world. His fame could even persuade the aged Taoist sage Ch'ang-ch'un to journey the length of Asia to discourse upon religious matters. He was above all adaptable, a man who could learn.

Organization, discipline, mobility, and ruthlessness of purpose were the fundamental factors in his military successes. Massacres of defeated populations, with the resultant terror, were weapons he regularly used. His practice of summoning cities to surrender and of organizing the methodical slaughter of those who did not submit has been described as psychological warfare; but, although it was undoubtedly policy to sap resistance by fostering terror, massacre was used for its own sake. Mongol practice, especially in the war against Khwārezm, was to send agents to demoralize and divide the garrison and populace of an enemy city, mixing threats with promises. The Mongols' reputation for frightfulness often paralyzed their captives, who allowed themselves to be killed when resistance or flight was not impossible. Indeed, the Mongols were unaccountable. Resistance brought certain destruction, but at Balkh, now in Afghanistan, the population was slaughtered in spite of a prompt surrender, for tactical reasons.

The achievements of Genghis Khan were grandiose. He united all the nomadic tribes, and with numerically infcrior armies he defeated great empires, such as Khwārezm and the even more powerful Chin state. Yet he did not exhaust his people. He chose his successor, his son Ögödei, with great care, ensured that his other sons would obey Ögödei, and passed on to him an army and a state in full vigour. At the time of his death, Genghis Khan had conquered the land mass extending from Peking to the Caspian Sea, and his generals had raided Persia and Russia. His successors would extend their power over the whole of China, Persia, and most of Russia. Thcy did what he did not achieve and perhaps ncver really intendcd—that is, to wcld their conquests into a tightly organized empire. The dcstruction brought about by Genghis Khan survives in popular memory, but far more significant, thesc conquests were but the first stage of the Mongol Empire, the grcatcst continental empire of medieval and modern times. (C.R.B.)

BIBLIOGRAPHY. PAUL RATCHNEVSKY, *Genghis Khan: His Life and Legacy* (1991; originally published in German, 1983), is an important, scholarly biography. Other biographies include B.I. VLADIMIRTSOV, *The Life of Chingis-Khan* (1930, reissued 1969; originally published in Russian, 1922); R.P. LISTER, *Genghis Khan* (1969, reissued 1989), an account of his early life based on a Mongol chronicle written in 1240 and recovered in the 20th century; LEO DE HARTOG, *Genghis Khan, Conqueror of the World* (1989; originally published in Dutch, 1979); and MICHEL HOANG, *Genghis Khan* (1990; originally published in French, 1988), emphasizing the military strategy and foresight of the ruler.

Accessible versions of the only contemporary native Mongolian portrait of the early imperial period may be found in FRANCIS WOODMAN CLEAVES (trans. and ed.), *The Secret History of the Mongols,* trans. from Mongolian (1982); and *The History and the Life of Chinggis Khan: The Secret History of the Mongols,* trans. from Mongolian and annotated by URGUNGE ONON (1990). For general reading, LEONARDO OLSCHKI, *Marco Polo's Asia* (1960; originally published in Italian, 1957), is recommended. EUSTACE D. PHILLIPS, *The Mongols* (1969), is a brief general history of the early Mongols, concentrating on the period of expansion in the 13th century. A carefully researched history of Mongol conquests, illustrated with maps and tables, is offered in PETER BRENT, *Genghis Khan* (also published as *The Mongol Empire,* 1976). DAVID MORGAN, *The Mongols* (1986), situates the achievements of Genghis Khan within a general survey of imperial Mongolia. A well-documented account of Mongol successes in Europe with specific military detail is found in JAMES CHAMBERS, *The Devil's Horsemen,* rev. and extended ed. (1988). (C.R.B./Ed.)

Geochronology: The Interpretation and Dating of the Geologic Record

Geochronology is a field of scientific investigation concerned with determining the age and history of rocks and rock assemblages. Such time determinations are made and the record of past geologic events is deciphered by studying the distribution and succession of rock strata, as well as the character of the fossil organisms preserved within the strata.

The Earth's surface is a complex mosaic of exposures of different rock types that are assembled in an astonishing array of geometries and sequences. Individual rocks in the myriad of rock outcroppings (or in some instances shallow subsurface occurrences) contain certain materials or mineralogic information that can provide insight as to their "age."

For years investigators determined the relative ages of sedimentary rock strata on the basis of their positions in an outcrop and their fossil content. According to a long-standing principle of the geosciences, that of superposition, the oldest layer within a sequence of strata is at the base and the layers are progressively younger with ascending order. The relative ages of the rock strata deduced in this manner can be corroborated and at times refined by the examination of the fossil forms present. The tracing and matching of the fossil content of separate rock outcrops (*i.e.*, correlation) eventually enabled investigators to integrate rock sequences in many areas of the world and construct a relative geologic time scale.

Scientific knowledge of the Earth's geologic history has advanced significantly since the development of radiometric dating, a method of age determination based on the principle that radioactive atoms in geologic materials decay at constant, known rates to daughter atoms. Radiometric dating has provided not only a means of numerically quantifying geologic time but also a tool for determining the age of various rocks that predate the appearance of life-forms.

This article traces the notable developments that gave rise to geochronology and describes the principal methods of relative and absolute dating devised over the years. It also surveys the major intervals of Earth history, delineating the rock systems, environmental features, and life-forms associated with each.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 241, 242, 243, 312, and 313, and the *Index*.

This article is divided into the following sections:

STUDY OF THE ROCK RECORD

Early views and discoveries

Some estimates suggest that as much as 70 percent of all rocks outcropping from the Earth's surface are sedimentary. Preserved in these rocks is the complex record of the many transgressions and regressions of the sea, as well as the fossil remains or other indications of now extinct organisms and the petrified sands and gravels of ancient beaches, sand dunes, and rivers.

Modern scientific understanding of the complicated story told by the rock record is rooted in the long history of observations and interpretations of natural phenomena extending back to the early Greek scholars. Xenophanes of Colophon (560?–478? BC), for one, saw no difficulty in describing the various seashells and images of life-forms embedded in rocks as the remains of long-deceased organisms. In the correct spirit but for the wrong reasons, Herodotus (5th century BC) felt that the small discoidal nummulitic petrifactions (actually the fossils of ancient lime-secreting marine protozoans) found in limestones outcropping at al-Jizah, Egypt, were the preserved remains of discarded lentils left behind by the builders of the pyramids.

These early observations and interpretations represent the unstated origins of what was later to become a basic principle of uniformitarianism, the root of any attempt at linking the past (as preserved in the rock record) to the present. Loosely stated, the principle says that the various natural phenomena observed today must also have existed in the past (see below *The emergence of modern geologic thought: Lyell's promulgation of uniformitarianism*).

Although quite varied opinions about the history and origins of life and of the Earth itself existed in the pre-Christian era, a divergence between Western and Eastern thought on the subject of natural history became more pronounced as a result of the extension of Christian dogma to the explanation of natural phenomena. Increasing constraints were placed upon the interpretation of nature in view of the teachings of the Bible. This required that the Earth be conceived of as a static, unchanging body, with a history that began in the not too distant past, perhaps as little as 6,000 years earlier, and an end, according to the scriptures, that was in the not too distant future. This biblical history of the Earth left little room for interpreting the Earth as a dynamic, changing system. Past catastrophes, particularly those that may have been responsible for altering the Earth's surface such as the great flood of Noah, were considered an artifact of the earliest formative history of the Earth. As such, they were considered unlikely to recur on what was thought to be an unchanging world.

With the exception of a few prescient individuals such as Roger Bacon (*c.* 1220–92) and Leonardo da Vinci (1452–1519), no one stepped forward to champion an enlightened view of the natural history of the Earth until the mid-17th century. Leonardo seems to have been among the first of the Renaissance scholars to "rediscover" the uniformitarian dogma through his observations of fossil marine organisms and sediments exposed in the hills of northern Italy. He recognized that the marine organisms now found as fossils in rocks exposed in the Tuscan Hills were simply ancient animals that lived in the region when it had been covered by the sea and were eventually buried by muds along the seafloor. He also recognized that the rivers of northern Italy, flowing south from the Alps and emptying into the sea, had done so for a very long time.

In spite of this deductive approach to interpreting natural events and the possibility that they might be preserved and later observed as part of a rock outcropping, little or no attention was given to the history—namely, the sequence of events in their natural progression—that might be preserved in these same rocks.

THE PRINCIPLE OF SUPERPOSITION OF ROCK STRATA

In 1669 the Danish-born natural scientist Nicolaus Steno (née Niels Steensen) published his noted treatise *De solido intra solidum naturaliter contento dissertationis prodromus* (Eng. trans. *The Prodromus of Nicolaus Steno's Dissertation Concerning a Solid Body Enclosed by Process of Nature Within a Solid*). This seminal work laid the essential framework for the science of geology by showing in very simple fashion that the layered rocks of Tuscany exhibit sequential change—that they contain a record of past events. Following from this observation, Steno concluded that the Tuscan rocks demonstrated superpositional relationships: rocks deposited first lie at the bottom of a sequence, while those deposited later are at the top. This is the crux of what is now known as the principle of superposition. Steno put forth still another idea—that layered rocks were likely to be deposited horizontally. Therefore, even though the strata of Tuscany were (and still are) displayed in anything but simple geometries, Steno's elucidation of these fundamental principles relating to the formation of stratified rock made it possible to work out not only superpositional relationships within rock sequences but also the relative age of each layer.

Steno's contributions

With the publication of the *Prodromus* and the ensuing widespread dissemination of Steno's ideas, other natural scientists of the latter part of the 17th and early 18th centuries applied them to their own work. The early English geologist John Strachey, for example, produced in 1725 what may well have been the first modern geologic maps of rock strata. He also described the succession of strata associated with coal-bearing sedimentary rocks in Somersetshire, the same region of England where he had mapped the rock exposures.

CLASSIFICATION OF STRATIFIED ROCKS

In 1756 Johann Gottlob Lehmann of Germany reported on the succession of rocks in the southern part of his country and the Alps, measuring and describing their compositional and spatial variation. While making use of Steno's principle of superposition, Lehmann recognized the existence of three distinct rock assemblages: (1) a successionally lowest category, the Primary (Urgebirge), composed mainly of crystalline rocks, (2) an intermediate category, or the Secondary (Flötzgebirge), composed of layered or stratified rocks containing fossils, and (3) a final or successionally youngest sequence of alluvial and related unconsolidated sediments (Angeschwemmtgebirge) thought to represent the most recent record of the Earth's history.

This threefold classification scheme was successfully applied with minor alterations to studies in other areas of Europe by three of Lehmann's contemporaries. In Italy, again in the Tuscan Hills in the vicinity of Florence, Giovanni Arduino, regarded by many as the father of Italian geology, proposed a four-component rock succession. His Primary and Secondary divisions are roughly similar to Lehmann's Primary and Secondary categories. In addition, Arduino proposed another category, the Tertiary division, to account for poorly consolidated though stratified fossil-bearing rocks that were superpositionally older than the (overlying) alluvium but distinct and separate from the hard (underlying) stratified rocks of the Secondary.

In two separate publications, one that appeared in 1762 and the second in 1773, Georg Christian Füchsel also applied Lehmann's earlier concepts of superposition to another sequence of stratified rocks in southern Germany. While using upwards of nine separate categories of sedimentary rocks, Füchsel essentially identified discrete rock bodies of unique composition, lateral extent, and position within a rock succession. (These rock bodies would constitute formations in modern terminology.)

Nearly 1,000 kilometres (620 miles) to the east, the German naturalist Peter Simon Pallas was studying rock sequences exposed in the southern Urals of eastern Russia. His report of 1777 differentiated a threefold division of rock, essentially reiterating Lehmann's work by extension.

Thus, by the latter part of the 18th century, the superpositional concept of rock strata had been firmly established through a number of independent investigations throughout Europe. Although Steno's principles were being widely applied, there remained to be answered a number of fundamental questions relating to the temporal and lateral relationships that seemed to exist among these disparate European sites. Were these various German, Italian, and Russian sites at which Lehmann's threefold rock succession was recognized contemporary? Did they record the same series of geologic events in the Earth's past? Were the various layers at each site similar to those of other sites? In short, was correlation among these various sites now possible?

The emergence of modern geologic thought

Inherent in many of the assumptions underlying the early attempts at interpreting natural phenomena in the latter part of the 18th century was the ongoing controversy between the biblical view of Earth processes and history and a more direct approach based on what could be observed and understood from various physical relationships demonstrable in nature. A substantial amount of information about the compositional character of many rock sequences was beginning to accumulate at this time. Abraham Gottlob Werner, a scholar of wide repute and following from the School of Mining in Freiberg, Ger., was very successful in reaching a compromise between what could be said to be scientific "observation" and biblical "fact." Werner's theory was that all rocks (including the sequences being identified in various parts of Europe at that time) and the Earth's topography were the direct result of either of two processes: (1) deposition in the primeval ocean, represented by the Noachian flood (his two "Universal," or Primary, rock series), or (2) sculpturing and deposition during the retreat of this ocean from the land (his two "Partial," or disintegrated, rock series). Werner's interpretation, which came to represent the so-called Neptunist conception of the Earth's beginnings, found widespread and nearly universal acceptance owing in large part to its theological appeal and to Werner's own personal charisma.

Werner's Neptunist theory

One result of Werner's approach to rock classification was that each unique lithology in a succession implied its own unique time of formation during the Noachian flood and a universal distribution. As more and more comparisons were made of diverse rock outcroppings, it began to become apparent that Werner's interpretation did not "universally" apply. Thus arose an increasingly vocal challenge to the Neptunist theory.

James Hutton's recognition of the geologic cycle

Attack on Neptunism and catastrophism

In the late 1780s the Scottish scientist James Hutton launched an attack on much of the geologic dogma that had its basis in either Werner's Neptunist approach or its corollary that the prevailing configuration of the Earth's surface is largely the result of past catastrophic events which have no modern counterparts. Perhaps the quintessential spokesman for the application of the scientific method in solving problems presented in the complex world of natural history, Hutton took issue with the catastrophist and Neptunist approach to interpreting rock histories and instead used deductive reasoning to explain what he saw. By Hutton's account, the Earth could not be viewed as a simple, static world not currently undergoing change. Ample evidence from Hutton's Scotland provided the key to unraveling the often thought but still rarely stated premise that events occurring today at the Earth's surface—namely erosion, transportation and deposition of sediments, and volcanism—seem to have their counterparts preserved in the rocks. The rocks of the Scottish coast and the area around Edinburgh proved the catalyst for his argument that the Earth is indeed a dynamic, ever-changing system, subject to a sequence of recurrent cycles of erosion and deposition and of subsidence and uplift. Hutton's formulation of the principle of uniformitarianism, which holds that Earth processes occurring today had their counterparts in the ancient past, while not the first time that this general concept was articulated, was probably the most important geologic concept developed out of rational scientific thought of the 18th century. The publication of Hutton's two-volume *Theory of the Earth* in 1795 firmly established him as one of the founders of modern geologic thought.

It was not easy for Hutton to popularize his ideas, however. The *Theory of the Earth* certainly did set the fundamental principles of geology on a firm basis, and several of Hutton's colleagues, notably John Playfair with his *Illustrations of the Huttonian Theory of the Earth* (1802), attempted to counter the entrenched Wernerian influence of the time. Nonetheless, another 30 years were to pass before Neptunist and catastrophist views of Earth history were finally replaced by those grounded in a uniformitarian approach.

This gradual unseating of the Neptunist theory resulted from the accumulated evidence that increasingly called into question the applicability of Werner's Universal and Partial formations in describing various rock successions. Clearly, not all assignable rock types would fit into Werner's categories, either superpositionally in some local succession or as a unique occurrence at a given site. Also, it was becoming increasingly difficult to accept certain assertions of Werner that some rock types (*e.g.,* basalt) are chemical precipitates from the primordial ocean. It was this latter observation that finally rendered the Neptunist theory unsustainable. Hutton observed that basaltic rocks exposed in the Salisbury Craigs, just on the outskirts of Edinburgh, seemed to have baked adjacent enclosing sediments lying both below and above the basalt. This simple observation indicated that the basalt was emplaced within the sedimentary succession while it was still sufficiently hot to have altered the sedimentary material. Clearly, basalt could not form in this way as a precipitate from the primordial ocean as Werner had claimed. Furthermore, the observations at Edinburgh indicated that the basalt intruded the sediments from below—in short, it came from the Earth's interior, a process in clear conflict with Neptunist theory.

While explaining that basalt may be intrusive, the Salisbury Craigs observations did not fully satisfy the argument that some basalts are not intrusive. Perhaps the Neptunist approach had some validity? The resolution of this latter problem occurred at an area of recent volcanism in the Auvergne area of central France. Here, numerous cinder cones and fresh lava flows composed of basalt provided ample evidence that this rock type is the solidified remnant of material ejected from the Earth's interior, not a precipitate from the primordial ocean.

Lyell's promulgation of uniformitarianism

Hutton's words were not lost on the entire scientific community. Charles Lyell, another Scottish geologist, was a principal proponent of Hutton's approach, emphasizing gradual change by means of known geologic processes. In his own observations on rock and faunal successions, Lyell was able to demonstrate the validity of Hutton's doctrine of uniformitarianism and its importance as one of the fundamental philosophies of the geologic sciences. Lyell, however, imposed some conditions on uniformitarianism that perhaps had not been intended by Hutton: he took a literal approach to interpreting the principle of uniformity in nature by assuming that all past events must have conformed to controls exerted by processes that behaved in the same manner as those processes behave today. No accommodation was made for past conditions that do not have modern counterparts. In short, volcanic eruptions, earthquakes, and other violent geologic events may indeed have occurred earlier in Earth history but no more frequently nor with greater intensity than today; accordingly, the surface features of the Earth are altered very gradually by a series of small changes rather than by occasional cataclysmic phenomena.

Uniformity of causes and uniformity of intensity

Lyell's contribution enabled the doctrine of uniformitarianism to finally hold sway, even though it did impose for the time being a somewhat limiting condition on the uniformity principle. This, along with the increased recognition of the utility of fossils in interpreting rock successions, made it possible to begin addressing the question of the meaning of time in Earth history.

Determining the relationships of fossils with rock strata

The hypothesis of fossil succession in the work of Georges Cuvier. During this period of confrontation between the proponents of Neptunism and uniformitarianism, there emerged evidence resulting from a lengthy and detailed study of the fossiliferous strata of the Paris Basin that rock successions were not necessarily complete records of past geologic events. In fact, significant breaks frequently occur in the superpositional record. These breaks affect not only the lithologic character of the succession but also the character of the fossils found in the various strata.

An 1812 study by the French zoologist Georges Cuvier was prescient in its recognition that fossils do in fact record events in Earth history and serve as more than just "follies" of nature. Cuvier's thesis, based on his analysis of the marine invertebrate and terrestrial vertebrate fauna of the Paris Basin, showed conclusively that many fossils, particularly those of terrestrial vertebrates, had no living counterparts. Indeed, they seemed to represent extinct forms, which, when viewed in the context of the succession of strata with which they were associated, constituted part of a record of biological succession punctuated by numerous extinctions. These, in turn, were followed by a seeming renewal of more advanced but related forms and were separated from each other by breaks in the associated rock record. Many of these breaks were characterized by coarser, even conglomeratic strata following a break, suggesting "catastrophic" events that may have contributed to the extinction of the biota. Whatever the actual cause, Cuvier felt that the evidence provided by the record of faunal succession in the Paris Basin could be interpreted by invoking recurring catastrophic geologic events, which in turn contributed to recurring massive faunal extinction, followed at a later time by biological renewal.

William Smith's work with faunal sequence. As Cuvier's theory of faunal succession was being considered, William Smith, a civil engineer from the south of England, was also coming to realize that certain fossils can be found consistently associated with certain strata. In the course of evaluating various natural rock outcroppings, quarries, canals, and mines during the early 1790s, Smith increasingly utilized the fossil content as well as the lithologic character of various rock strata to identify the successional position of different rocks, and he made use of this information to effect a correlation among various localities he had studied. The consistency of the relationships that Smith observed eventually led him to conclude that there is indeed faunal succession and that there appears to be a consistent progression of forms from more primitive to more advanced. As a result of this observation, Smith was able to begin what was to amount to a monumental effort at synthesizing all that was then known of the rock successions outcropping throughout parts of Great Britain. This effort culminated in the publication of his "Geologic Map of England, Wales and Part of Scotland" (1815), a rigorous treatment of diverse geologic information resulting from a thorough understanding of geologic principles, including those of original horizontality, superposition (lithologic, or rock, succession), and faunal succession. With this, it now became possible to assume within a reasonable degree of certainty that correlation could be made between and among widely separated areas. It also became apparent that many sites that had previously been classified according to the then-traditional views of Arduino, Füchsel, and Lehmann did not conform to the new successional concepts of Smith.

Foundations of stratigraphic correlation

EARLY ATTEMPTS AT MAPPING AND CORRELATION

The seminal work of Smith at clarifying various relationships in the interpretation of rock successions and their correlations elsewhere resulted in an intensive look at what the rock record and, in particular, what the fossil record had to say about past events in the long history of the Earth. A testimony to Smith's efforts in producing one of the first large-scale geologic maps of a region is its essential accuracy in portraying what is now known to be the geologic succession for the particular area of Britain covered.

The application of the ideas of Lyell, Smith, Hutton, and others led to the recognition of lithologic and paleontologic successions of similar character from widely scattered areas. It also gave rise to the realization that many of these similar sequences could be correlated.

The French biologist Jean-Baptiste de Monet, Chevalier de Lamarck, in particular, was able to demonstrate the similarity of fauna from a number of Cuvier's and Alexandre Brongniart's collections of fossils from the Paris Basin with fossil fauna from the sub-Apennines of Italy and the London Basin. While based mainly on the collections of Cuvier and Brongniart, Lamarck's observations provided much more insight into the real significance of using fossils strictly for correlation purposes. Lamarck disagreed with Cuvier's interpretation of the meaning of faunal extinction and regeneration in stratigraphic successions. Not convinced that catastrophes caused massive and widespread disruption of the biota, Lamarck preferred to think of organisms and their distribution in time and space as responding to the distribution of favourable habitats. If confronted with the need to adapt to abrupt changes in local habitat—Cuvier's catastrophes—faunas must be able to change in order to survive. If not, they became extinct. Lamarck's approach, much like that of Hutton, stressed the continuity of processes and the continuum of the stratigraphic record. Moreover, his view that organisms respond to the conditions of their environment had important implications for the uniformitarian approach to interpreting Earth history.

The impact of Lamarck's work

Once it was recognized that many of the rocks of the Paris Basin, London Basin, and parts of the Apennines apparently belonged to the same sequence by virtue of the similarity of their fossil content, Arduino's term Tertiary (proposed as part of his fourfold division of rock succession in the Tuscan Hills of Italy) began to be applied to all of these diverse locations. Further work by Lyell and Gérard-Paul Deshayes resulted in the term Tertiary being accepted as one of the fundamental divisions of geologic time.

THE CONCEPTS OF FACIES, STAGES, AND ZONES

Facies. During the latter half of the 18th and early 19th centuries, most of the research on the distribution of rock strata and their fossil content treated lithologic boundaries as events in time representing limits to strata that contain unique lithology and perhaps a unique fossil fauna, all of which are the result of unique geologic processes acting over a relatively brief period of time. Hutton recognized early on, however, that some variations occur in the sediments and fossils of a given stratigraphic unit and that such variations might be related to differences in depositional environments. He noted that processes such as erosion in the mountains of Scotland, transportation of sand and gravels in streams flowing from these mountains, and the deposition of these sediments could all be observed to be occurring concurrently. At a given time then, these diverse processes were all taking place at separate locations. As a consequence, different environments produce different sedimentary products and may harbour different organisms. This aspect of differing lithologic type or environmental or biological condition came to be known as facies. (It was Steno who had, in 1669, first used the term facies in reference to the condition or character of the Earth's surface at a particular time.)

The significance of the facies concept for the analysis of geologic history became fully apparent with the findings of the Swiss geologist Amanz Gressly. While conducting survey work in the Jura Mountains in 1838, Gressly observed that rocks from a given position in a local stratigraphic succession frequently changed character as he traced them laterally. He attributed this lateral variation to lateral changes in the depositional environments responsible for producing the strata in question. Having no term to apply to the observed changes, he adopted the word facies. While Gressly employed the term specifically in the context of lithologic character, it is applied more broadly today. As now used, the facies concept has come to encompass other types of variation that may be encountered as one moves laterally (*e.g.,* along outcroppings of rock strata exposed in stream valleys or mountain ridges) in a given rock succession. Lithologic facies, biological facies, and even environmental facies can be used to describe sequences of rocks of the same or different age having a particularly unique character.

Modern usage of the facies concept

Stages and zones. The extensive review of the marine invertebrate fauna of the Paris Basin by Deshayes and Lyell not only made possible the formalization of the term Tertiary but also had a more far-reaching effect. The thousands of marine invertebrate fossils studied by Deshayes enabled Lyell to develop a number of subdivisions of the Tertiary of the Paris Basin based on the quantification of molluskan species count and duration. Lyell noted that of the various assemblages of marine mollusks found, those from rocks at the top of the succession contained a large number of species that were still extant in modern environments. Progressively older strata yielded fewer and fewer forms that had living counterparts, until at the base of the succession, a very small number of the total species present could be recognized as having modern counterparts. This fact allowed Lyell to consider subdividing the Tertiary of the Paris Basin into smaller increments, each of which could be defined according to some relative percentage of living species present in the strata. The subdivision resulted in the delineation of the Eocene, Miocene, and Pliocene epochs in 1833. Later this scheme was refined to further divide the Pliocene into an Early and a Late Pliocene.

Lyell's biostratigraphically defined concept of sequence, firmly rooted in concepts of faunal succession and superposition, was developed on mixed but stratigraphically controlled collections of fossils. It worked, but it did not address the faunal composition of the various Paris Basin strata other than in gross intervals—intervals that were as much lithologically as paleontologically defined. Alcide d'Orbigny, a French geologist, demonstrated correlational and superpositional uniqueness by utilizing paleontologically distinct intervals of strata defined solely on the basis of their fossil assemblages in his study of the French Jurassic *Terrains Jurassiques* (1842). This departure from a lithologically based concept of paleontologic succession enabled d'Orbigny to define paleontologically unique stages. Each stage represented a unique period in time and formed the basis of later work that resulted in the further subdivision of d'Orbigny's original stages into 10 distinct stage assemblages. In spite of the work of Smith and to a lesser extent Lyell and others, d'Orbigny's approach was essentially that of a catastrophist. Stage boundaries were construed to represent unusual extrinsic geologic events, with significant implications for faunal continuity. The applicability of d'Orbigny's stages to areas outside of France had only limited success. At this point in the development of paleontology as a science, little was understood about the geologic time range of various fauna. Even less was known about the habitats—the environmental limits—of ancient fauna. Could certain groups of organisms have sufficiently widespread distribution in the rock record to enable correlations to be made with certainty? The Jurassic of western Europe consisted mostly of shallow marine sediments widely deposited throughout the area. It is now known that some of the mollusks with which d'Orbigny worked were undergoing very rapid evolutionary change; they were thus relatively short-lived as distinct forms in the geologic record and had a wide-ranging environmental tolerance. The result was that some forms, notably of the group of mollusks called ammonite cephalopods, were distributed extensively within a variety of sedimentary facies. The correlating of strata based on the faunal stage approach was widely accepted. Interestingly, most of d'Orbigny's Jurassic stages, with refinements, are still in use today.

Paleontologically distinct intervals of strata

Only a short time after d'Orbigny's original analysis of Jurassic strata, the German mineralogist and paleontologist Friedrich A. Quenstedt challenged (in 1856–58) the validity of using stages to effect correlations in cases where the actual geologic ranges and bed-by-bed distribution of individual component fossils of an assemblage were unknown. In retrospect, this seems blatantly obvious, but at the time the systematic stratigraphic documentation of fossil occurrence was not always carried out. Much critical biostratigraphic data necessary for the proper characterization of faunal assemblages was simply not collected. As argued, individual fossil ranges and their distributions could have profound influence on the concept of faunal succession and evolutionary dynamics.

Oppel's refinement of biostratigraphic concepts

Several of Quenstedt's students at the University of Tübingen followed up on this latter concern. One in particular, Carl Albert Oppel, essentially refined his mentor's concepts by paying particular attention to the character of the range of individual species in a succession of fauna. These intervals of unique biological character, which he called zones, were essentially subdivisions of the stages proposed by Quenstedt. Oppel's recognition of the earliest occurrence of a fossil species (or its first appearance), its range through a succession of strata, and its eventual loss from the local record (or its last appearance) led him to compare such biostratigraphic data from many species. By making use of such data on species that overlap in some or all of their stratigraphic ranges and from widely separated areas, Oppel was able to erect a biochronology based on a diverse record of first appearances, last appearances, and individual and overlapping range zones. This fine-scale refinement of a biologically defined sense of succession found wide applicability and enabled not only biochronological (or temporal) but also biofacies (spatial) understanding of the succession in question.

Figure 1 illustrates six of the Jurassic zones and some of the fossil species on which they are based. The zone of *Arietites bucklandi* is based on the joint occurrences of two species with differing ranges in time, here called "species J" and "G," of the *Arnioceras semicostatum* zone by overlap in the range of "species G" and "Q," and so on. The zonal fossils are characteristic but are not necessarily confined in their occurrence to the zone. (V.S.M./G.D.J.)

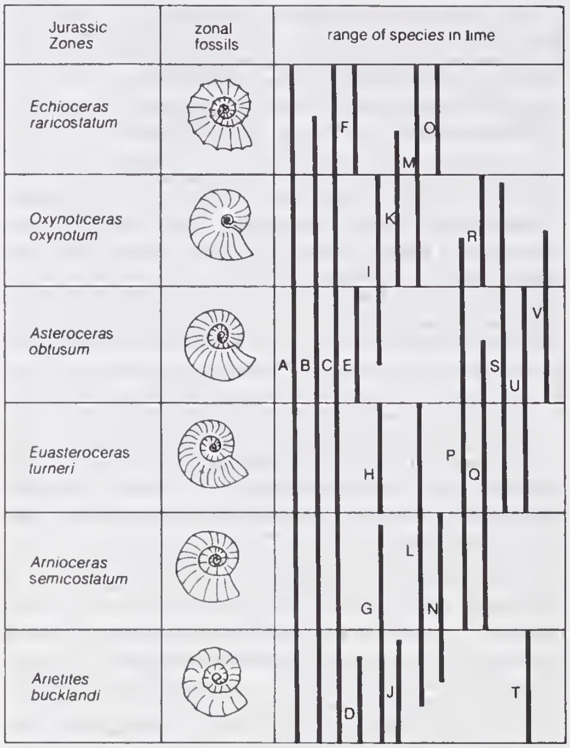


Figure 1: *Jurassic fossil zones and zonal fossils.*
These six ammonite species never overlap in stratigraphic distribution and therefore provide a guide for correlation wherever they occur. The ranges of hypothetical species (indicated by letters) within the Jurassic are also shown.

Based on Arkell in James R Beerbower, *Search for the Past. An Introduction to Paleontology*, 2nd ed., © 1968, by permission of Prentice-Hall, Inc

Completion of the Phanerozoic time scale

With the development of the basic principles of faunal succession and correlation and the recognition of facies variability, it was a relatively short step before large areas of Europe began to be placed in the context of a global geologic succession. This was not, however, accomplished in a systematic manner. Whereas the historical ideas of Lehmann and Arduino were generally accepted, it became increasingly clear that many diverse locally defined rock successions existed, each with its own unique fauna and apparent position within some sort of "universal" succession.

As discussed above, Arduino's Tertiary was recognized in certain areas and was in fairly common use after 1760, but only rudimentary knowledge of other rock successions existed by the later part of the 18th century. The German naturalist Alexander von Humboldt had recognized the widespread occurrence of fossil-bearing limestones throughout Europe. Particular to these limestones, which formed large tracts of the Jura Mountains of Switzerland, were certain fossils that closely resembled those known from the Lias and Oolite formations of England, which were then being described by William Smith. Subsequently, Humboldt's "Jura Kalkstein" succession, as he described it in 1795, came to be recognized throughout Europe and England. By 1839, when the geologist Leopold Buch recognized this rock sequence in southern Germany, the conceptual development of the Jurassic System was complete.

The coal-bearing strata of England, known as the Coal Measures, had been exploited for centuries, and their distribution and vertical and lateral variability were the subject of numerous local studies throughout the 17th and early 18th centuries, including those of Smith. In 1808 the geologist Jean-Baptiste-Julien d'Omalius d'Halloy described a coal-bearing sequence in Belgium as belonging to the Ter-

Refinement of systems

rain Bituminifère. Although the name did not remain in common usage for long, the Terrain Bituminifère found analogous application in the work of two English geologists, William D. Conybeare and William Phillips, in their synthesis of the geology of England and Wales in 1822. Conybeare and Phillips coined the term Carboniferous (or coal-bearing) to apply to the succession of rocks from north-central England that contained the Coal Measures. The unit also included several underlying rock formations extending down into what investigators now consider part of the underlying Devonian System. At the time, however, the approach by Conybeare and Phillips was to encompass in their definition of the Carboniferous all of the associated strata that could be reasonably included in the Coal Measures succession.

D'Omalius mapped and described a local succession in western France. While doing so, he began to recognize a common sequence of soft limestones, greensands (glauconite-bearing sandstones), and related marls in what is today known to be a widespread distribution along coastal regions bordering the North Sea and certain regions of the Baltic. The dominant lithology of this sequence is frequently the soft limestones or chalk beds so well known from the Dover region of southeast England and Calais in nearby France. D'Omalius called this marl, greensand, and chalk-bearing interval the Terrain Crétacé. Along with their adoption of the term Carboniferous in 1822, Conybeare and Phillips referred to the French Terrain Crétacé as the Cretaceous System.

Clearly, surficial deposits and related unconsolidated material, variously relegated to the categories of classification proposed by Arduino, Lehmann, Werner, and others as "alluvium" or related formations, deserved a place in any formalized system of rock succession. In 1829 Jules Desnoyers of France, studying sediments in the Seine valley, proposed using the term Quaternary to encompass all of these various post-Tertiary formations. At nearly the same time, the important work of Lyell on the faunal succession of the Paris Basin permitted finer-scaled discrimination of this classic Tertiary sequence. In 1833 Lyell, using various biostratigraphic evidence, proposed several divisions of the Tertiary System that included the Eocene, Miocene, and Pliocene epochs. By 1839 he proposed using the term Pleistocene instead of dividing his Pliocene Epoch into older and newer phases. The temporal subdivision of the Tertiary was completed by two German scientists, Heinrich Ernst Beyrich and Wilhelm Philipp Schimper. Beyrich introduced the Oligocene in 1854 after having investigated outcrops in Belgium and Germany, while Schimper proposed adding the Paleocene in 1874 based on his studies of Paris Basin flora.

Werner's quadripartite division of rocks in southern Germany was applied well into the second decade of the 19th century. During this time, rock sequences from the lower part of his third temporal subdivision, the Flötzgebirge, were subsequently subdivided into three formations, each having fairly widespread exposure and distribution. Based on his earlier work, Friedrich August von Alberti identified in 1834 these three distinct lithostratigraphic units, the Bunter Sandstone, the Muschelkalk Limestone, and the Keuper Marls and Clays, as constituting the Trias or Triassic System.

The Sedgwick–Murchison dispute

Perhaps one of the most intriguing episodes in the development of the geologic time scale concerns the efforts of two British geologists and in large measure their attempts at unraveling the complex geologic history of Wales. Adam Sedgwick and Roderick Impey Murchison began working, in 1831, on the sequence of rocks lying beneath the Old Red Sandstone (which had been included in the basal sequence of the Carboniferous, as defined by Conybeare and Phillips, earlier in 1822). What started as an earnest collaborative attempt at deciphering the structurally and stratigraphically complicated rock succession in Wales ended in 1835 with a presentation outlining two distinct subdivisions of the pre-Carboniferous succession. Working up from the base of the post-Primary rock succession of poorly fossiliferous clastic rocks in northern Wales, Sedgwick identified a sequence of rock units defined primarily by their various lithologies. He designated this succession the Cambrian, after Cambria, the Roman name for Wales. Murchison worked downward in the considerably more fossiliferous pre-Old Red Sandstone rock sequence in southern Wales and was able to identify a succession of strata containing a well-preserved fossil fauna. These sequences defined from southern Wales were eventually brought into the context of Sedgwick's Cambrian. Murchison named his rock succession the Silurian, after the Roman name for an early Welsh tribe. In a relatively short time, Murchison's Silurian was expanding both laterally and temporally as more and more localities containing the characteristic Silurian fauna were recognized throughout Europe. The major problem created by this conceptual "expansion" of the Silurian was that it came to be recognized in northern Wales as coincident with much of the strata in the upper portion of Sedgwick's Cambrian. With Sedgwick's Cambrian based mainly on lithologic criteria, the presence of Silurian fauna created correlational difficulties. As it turned out, Sedgwick's Cambrian was of little value outside of its area of original definition. With it being superseded by the paleontologically based concept of the Silurian, some sort of compromise had to be worked out.

This compromise came about primarily as a result of the work of Charles Lapworth, the English geologist who in 1879 proposed the designation Ordovician System for that sequence of rocks representing the upper part of Sedgwick's Cambrian succession and the lower (and generally overlapping) portion of Murchison's Silurian succession. The term Ordovician is derived from yet another Roman-named tribe of ancient Wales, the Ordovices. A large part of Lapworth's rationale for this division was based on the earlier work of the French-born geologist Joachim Barrande, who investigated the apparent Silurian fauna of central Bohemia. Barrande's 1851 treatise on this area of Czechoslovakia demonstrated a distinct succession from a "second" Silurian fauna to a "third" Silurian fauna. This divisible Silurian, as well as separate lines of evidence gathered by Lapworth in Scotland and Wales, finally enabled the individual character of the Cambrian, Ordovician, and Silurian systems to be resolved.

While involved in their work on Welsh stratigraphic successions, Sedgwick and Murchison had the opportunity to compare some rock outcroppings in Devonshire, in southwest England, with similar rocks in Wales. The Devon rocks were originally thought to belong to part of Sedgwick's Cambrian System, but they contained plant fossils very similar to basal Carboniferous (Old Red Sandstone) plant fossils found elsewhere. Eventually recognizing that these fossil-bearing sequences represented lateral equivalents in time and perhaps temporally unique strata as well, Sedgwick and Murchison in 1839 proposed the Devonian System.

During the early 1840s, Murchison traveled with the French paleontologist Edouard de Verneuil and the Latvian-born geologist Alexandr Keyserling to study the rock succession of the eastern Russian platform, the area of Russia west of the Ural Mountains. Near the town of Perm, Murchison and Verneuil identified fossiliferous strata containing both Carboniferous and a younger fauna at that time not recognized elsewhere in Europe or in the British Isles. Whereas the Carboniferous fossils were similar to those they had seen elsewhere (mainly from the Coal Measures), the stratigraphically higher fauna appeared somewhat transitional to the Triassic succession of Germany as then understood. Murchison coined the term Permian (after the town of Perm) to represent this intermediate succession.

The Carboniferous System and its North American equivalents

With continued refinement of the definition of the Carboniferous in Europe, particularly in England, what at one time comprised the Old Red Sandstone, Lower Coal Measures (Mountain Limestone and Millstone Grit), and Upper Coal Measures now stood as just the Lower and Upper Coal Measures. It was beginning to be recognized that certain rock sequences in the Catskill Mountains of eastern New York state in North America resembled the Old Red Sandstone of western England. Furthermore, coal-bearing strata exposed in Pennsylvania greatly resembled the similar coal-bearing strata of the Upper Coal

Measures. Lying beneath these coal-bearing rocks of Pennsylvania was a sequence of limestones that could be traced over thousands of square kilometres and that occurred in numerous outcrops along various tributary streams to the Ohio and Mississippi rivers in Indiana, Kentucky, Missouri, Illinois, and Iowa. This "subcarboniferous" strata, identified by the American geologist David Dale Owen in 1839, was subsequently termed Mississippian in 1870 as a result of work conducted by another American geologist, Alexander Winchell, in the upper Mississippi valley area. Eventually the overlying strata, the coal-bearing rocks originally described from Pennsylvania, were formalized as Pennsylvanian in 1891 by the paleontologist and stratigrapher Henry Shaler Williams.

The North American-defined Mississippian and Pennsylvanian systems were later correlated with presumed European and British successions. Although approximately similar in successional relationship, the Mississippian–Pennsylvanian boundary in North America is now considered slightly younger than the Lower–Upper Carboniferous boundary in Europe.

By the 1850s, with the development of the geologic time scale nearly complete, investigators were beginning to recognize that a number of major paleontologically defined boundaries were common and recurrent regardless of where a succession was studied. By this time rock successions were being defined according to fauna they contained, and the relative time scale, which was being erected, was based on the principle of faunal succession; consequently, any major hiatus or change in faunal character was bound to be interpreted as important. In 1838 Sedgwick proposed that all pre-Old Red Sandstone sediments be included in the rock succession designated the Paleozoic Series (or Era) that contained generally primitive fossil fauna. John Phillips, another English geologist, went on to describe the Mesozoic Era to accommodate what then was the Cretaceous, Jurassic, Triassic, and partially Permian strata, and the Kainozoic (Cainozoic, or Cenozoic) era to include Lyell's Eocene, Miocene, and Pliocene. This subdivision of the generally fossiliferous strata that lay superpositionally above the so-called Primary rocks of many of the early workers resulted in the recognition of three distinct eras. Subsequent subdivision of these eras into specific geologic periods finally provided the hierarchy for describing the relative dating of geologic events.

Development of radioactive dating methods and their application

As has been seen, the geologic time scale is based on stratified rock assemblages that contain a fossil record. For the most part, these fossils allow various forms of information from the rock succession to be viewed in terms of their relative position in the sequence. Approximately the first 87 percent of Earth history occurred before the evolutionary development of shell-bearing organisms. The result of this mineralogic control on the preservability of organic remains in the rock record is that the geologic time scale—essentially a measure of biologic changes through time—takes in only the last 13 percent of Earth history. Although the span of time preceding the Cambrian period—the Precambrian—is nearly devoid of characteristic fossil remains and coincides with some of the primary rocks of certain early workers, it must, nevertheless, be evaluated in its temporal context.

EARLY ATTEMPTS AT CALCULATING THE AGE OF THE EARTH

Difficulties of dating Precambrian events

Historically, the subdivision of Precambrian rock sequences (and, therefore, Precambrian time) had been accomplished on the basis of structural or lithologic grounds. With only minor indications of fossil occurrence (mainly in the form of algal stromatolites), no effective method of quantifying this loosely constructed chronology existed until the discovery of radioactivity enabled dating procedures to be applied directly to the rocks in question.

The quantification of geologic time remained an elusive matter for most human enquiry into the age of the Earth and its complex physical and biological history. Although Hindu teachings accept a very ancient origin for the Earth, medieval Western concepts of Earth history were based for the most part on a literal interpretation of Old Testament references. Biblical scholars of Renaissance Europe and later considered paternity as a viable method by which the age of the Earth since its creation could be determined. A number of attempts at using the "begat" method of determining the antiquity of an event—essentially counting backward in time through each documented human generation—led to the age of the Earth being calculated at several thousand years. One such attempt was made by Archbishop James Ussher of Ireland, who in 1650 determined that the Creation had occurred during the evening of Oct. 22, 4004 BC. By his analysis of biblical genealogies, the Earth was not even 6,000 years old!

From the time of Hutton's refinement of uniformitarianism, the principle found wide application in various attempts to calculate the age of the Earth. As previously noted, fundamental to the principle was the premise that various Earth processes of the past operated in much the same way as those processes operate today. The corollary to this was that the rates of the various ancient processes could be considered the same as those of the present day. Therefore, it should be possible to calculate the age of the Earth on the basis of the accumulated record of some process that has occurred at this determinable rate since the Creation.

Many independent estimates of the age of the Earth have been proposed, each made using a different method of analysis. Some such estimates were based on assumptions concerning the rate at which dissolved salts or sediments are carried by rivers, supplied to the world's oceans, and allowed to accumulate over time. These chemical and physical arguments (or a combination of both) were all flawed to varying degrees because of an incomplete understanding of the processes involved. The notion that all of the salts dissolved in the oceans were the products of leaching from the land was first proposed by the English astronomer and mathematician Edmond Halley in 1691 and restated by the Irish geologist John Joly in 1899. It was assumed that the ocean was a closed system and that the salinity of the oceans was an ever-changing and ever-increasing condition. Based on these calculations, Joly proposed that the Earth had consolidated and that the oceans had been created between 80 and 90 million years ago. The subsequent recognition that the ocean is not closed and that a continual loss of salts occurs due to sedimentation in certain environments severely limited this novel approach.

Equally novel but similarly flawed was the assumption that, if a cumulative measure of all rock successions were compiled and known rates of sediment accumulation were considered, the amount of time elapsed could be calculated. While representing a reasonable approach to the problem, this procedure did not or could not take into account different accumulation rates associated with different environments or the fact that there are many breaks in the stratigraphic record. Even observations made on faunal succession proved that gaps in the record do occur. How long were these gaps? Do they represent periods of nondeposition or periods of deposition followed by periods of erosion? Clearly sufficient variability in a given stratigraphic record exists such that it may be virtually impossible to even come to an approximate estimate of the Earth's age based on this technique. Nevertheless, many attempts using this approach were made (see Table 1).

William Thomson (later Lord Kelvin) applied his thermodynamic principles to the problems of heat flow, and this had implications for predicting the age of a cooling Sun and of a cooling Earth. From an initial estimate of 100 million years for the development of a solid crust around a molten core proposed in 1862, Thomson subsequently revised his estimate of the age of the Earth downward. Using the same criteria, he concluded in 1899 that the Earth was between 20 and 40 million years old.

Thomson's calculation was based on the assumption that the substance of the Earth is inert and thus incapable of producing new heat. His estimate came into question after the discovery of naturally occurring radioactivity by the

Table 1: Estimates of the Age of the Earth

date	author	maximum thickness (feet)	rate of deposit (years for 1 foot)	age* (millions of years)
1860	Phillips	72,000	1,332	96
1869	Huxley	100,000	1,000	100
1871	Haughton	177,200	8,616	1,526
1878	Haughton	177,200	?	200
1883	Winchell	—	—	3
1889	Croll	12,000†	6,000‡	72
1890	de Lapparent	150,000	600	90
1892	Wallace	177,200	158	28
1892	Geikie	100,000	730–6,800	73–680
1893	McGee	264,000	6,000	1,584
1893	Upham	264,000	316	100
1893	Walcott	—	—	45–70
1893	Reade	31,680†	3,000‡	95
1895	Sollas	164,000	100	17
1897	Sederholm	—	—	35–40
1899	Geikie	—	—	100
1900	Sollas	265,000	100	26.5
1908	Joly	265,000	300	80
1909	Sollas	335,000	100	80

*Based on estimates of maximum thicknesses of sedimentary rocks.
†Spread evenly over the land areas. ‡Rate of denudation.
Source: From D.R. Prothero, *Interpreting the Stratigraphic Record.* Copyright © 1990 by W.H. Freeman and Company. Reprinted with permission.

French physicist Henri Becquerel in 1896 and the subsequent recognition by his colleagues, Marie and Pierre Curie, that compounds of radium (which occur in uranium minerals) produce heat. As a result of this and other findings, notably that of Ernest Rutherford (see below), it became apparent that naturally occurring radioactive elements in minerals common in the Earth's crust are sufficient to account for all observed heat flow. Within a short time another leading British physicist, John William Strutt, concluded that the production of heat in the Earth's interior was a dynamic process, one in which heat was continuously provided by such materials as uranium. The Earth was, in effect, not cooling.

AN ABSOLUTE AGE FRAMEWORK FOR THE STRATIGRAPHIC TIME SCALE

Disintegration theory

In his book *Radio-activity* (1904), Rutherford explained that radioactivity results from the spontaneous disintegration of an unstable element into a lighter element, which may decay further until a stable element is finally created. This process of radioactive decay involves the emission of positively charged particles (later to be recognized as helium nuclei) and negatively charged ones (electrons) and in most cases gamma rays (a form of electromagnetic radiation) as well. This interpretation, the so-called disintegration theory, came to provide the basis for the numerical quantification of geologic time. (For additional information on radioactive decay and the principles of radiometric dating, see *Absolute dating* below.)

In 1905 Strutt succeeded in analyzing the helium content of a radium-containing rock and determined its age to be 2 billion years. This was the first successful application of a radiometric technique to the study of Earth materials, and it set the stage for a more complete analysis of geologic time. Although faced with problems of helium loss and therefore not quite accurate results, a major scientific breakthrough had been accomplished. Also in 1905 the American chemist Bertram B. Boltwood, working with the more stable uranium–lead system, calculated the numerical ages of 43 minerals. His results, with a range of 400 million to 2.2 billion years, were an order of magnitude greater than those of the other "quantitative" techniques of the day that made use of heat flow or sedimentation rates to estimate time.

Acceptance of these new ages was slow in coming. Perhaps much to their relief, paleontologists now had sufficient time in which to accommodate faunal change. Researchers in other fields, however, were still conservatively sticking with ages on the order of several hundred million, but were revising their assumed sedimentation rates downward in order to make room for expanded time concepts.

In a brilliant contribution to resolving the controversy over the age of the Earth, Arthur Holmes, a student of Strutt, compared the relative (paleontologically determined) stratigraphic ages of certain specimens with their numerical ages as determined in the laboratory. This 1911 analysis provided for the first time the numerical ages for rocks from several Paleozoic geologic periods as well as from the Precambrian. Carboniferous-aged material was determined to be 340 million years, Devonian-aged material 370 million years, Ordovician (or Silurian) material 430 million years, and Precambrian specimens from 1.025 to 1.64 billion years. As a result of this work, the relative geologic time scale, which had taken nearly 200 years to evolve, could be numerically quantified. No longer did it have merely superpositional significance, it now had absolute temporal significance as well. (G.D.J.)

Nonradiometric dating

In addition to radioactive decay, many other processes have been investigated for their potential usefulness in absolute dating. Unfortunately, they all occur at rates that lack the universal consistency of radioactive decay. Sometimes human observation can be maintained long enough to measure present rates of change, but it is not at all certain on a priori grounds whether such rates are representative of the past. This is where radioactive methods frequently supply information that may serve to calibrate nonradioactive processes so that they become useful chronometers. Nonradioactive absolute chronometers may conveniently be classified in terms of the broad areas in which changes occur—namely, geologic and biological processes, which will be treated here.

GEOLOGIC PROCESSES AS ABSOLUTE CHRONOMETERS

Weathering processes. During the first third of the 20th century, several presently obsolete weathering chronometers were explored. Most famous was the attempt to estimate the duration of Pleistocene interglacial intervals through depths of soil development. In the American Midwest, thicknesses of gumbotil and carbonate-leached zones were measured in the glacial deposits (tills) laid down during each of the four glacial stages. Based on a direct proportion between thickness and time, the three interglacial intervals were determined to be longer than postglacial time by factors of 3, 6, and 8. To convert these relative factors into absolute ages required an estimate in years of the length of postglacial time. When certain evidence suggested 25,000 years to be an appropriate figure, factors became years—namely, 75,000, 150,000, and 200,000 years. And, if glacial time and nonglacial time are assumed approximately equal, the Pleistocene Epoch lasted about 1,000,000 years.

Obsidian hydration

Only one weathering chronometer is employed widely at the present time. Its record of time is the thin hydration layer at the surface of obsidian artifacts. Although no hydration layer appears on artifacts of the more common flint and chalcedony, obsidian is sufficiently widespread that the method has broad application.

In a specific environment the process of obsidian hydration is theoretically described by the equation $D = Kt^{1/2}$, in which D is thickness of the hydration rim, K is a constant characteristic of the environment, and t is the time since the surface examined was freshly exposed. This relationship is confirmed both by laboratory experiments at 100° C (212° F) and by rim measurements on obsidian artifacts found in carbon-14 dated sequences (see below *Carbon-14 dating and other cosmogenic methods*). Practical experience indicates that the constant K is almost totally dependent on temperature and that humidity is apparently of no significance. Whether in a dry Egyptian tomb or buried in wet tropical soil, a piece of obsidian seemingly has a surface that is saturated with a molecular film of water. Consequently, the key to absolute dating of obsidian is to evaluate K for different temperatures. Ages follow from the above equation provided there is accurate knowledge of a sample's temperature history. Even without such knowledge, hydration rims are useful for relative dating within a region of uniform climate.

Like most absolute chronometers, obsidian dating has its problems and limitations. Specimens that have been exposed to fire or to severe abrasion must be avoided. Furthermore, artifacts reused repeatedly do not give ages corresponding to the culture layer in which they were found but instead to an earlier time, when they were fashioned. Finally, there is the problem that layers may flake off beyond 40 micrometres (0.004 centimetre, or 0.002 inch) of thickness—*i.e.,* more than 50,000 years in age. Measuring several slices from the same specimen is wise in this regard, and such a procedure is recommended regardless of age.

Accumulational processes. Sediment in former or present water bodies, salt dissolved in the ocean, and fluorine in bones are three kinds of natural accumulations and possible time indicators. To serve as geochronometers, the records must be complete and the accumulation rates known.

The fossiliferous part of the geologic column includes perhaps 122,000 metres of sedimentary rock if maximum thicknesses are selected from throughout the world. During the late 1800s, attempts were made to estimate the time over which it formed by assuming an average rate of sedimentation. Because there was great diversity among the rates assumed, the range of estimates was also large—from a high of 2.4 billion years to a low of 3 million years. In spite of this tremendous spread, most geologists felt that time in the hundreds of millions of years was necessary to explain the sedimentary record.

Varve dating

If the geologic column (see below) were made up entirely of annual layers, its duration would be easy to determine. Limited sedimentary deposits did accumulate in this way, and they are said to be varved; one year's worth of sediment is called a varve, and, in general, it includes two laminae per year.

Varves arise in response to seasonal changes. New Mexico's Castile Formation, for example, consists of alternating layers of gypsum and calcite that may reflect an annual temperature cycle in the hypersaline water from which the minerals precipitated. In moist, temperate climates, lake sediments collecting in the summer are richer in organic matter than those that settle during winter. This feature is beautifully seen in the seasonal progression of plant microfossils found in shales at Oensingen, Switz. In the thick oil shales of Wyoming and Colorado in the United States, the flora is not so well defined, but layers alternating in organic richness seem to communicate the same seasonal cycle. These so-called Green River Shales also contain abundant freshwater-fish fossils that confirm deposition in a lake. At their thickest, they span 792 vertical metres. Because the average thickness of a varve is about 0.015 centimetre (0.006 inch), the lake is thought to have existed for more than 5 million years.

Each of the examples cited above is of a floating chronology—*i.e.,* a decipherable record of time that was terminated long ago. In Sweden, by contrast, it has been possible to tie a glacial varve chronology to present time, and so create a truly absolute dating technique. Where comparisons with radiocarbon dating are possible, there is general agreement.

Fluorine dating

As early as 1844, an English chemist named Middleton claimed that fossil bones contain fluorine in proportion to their antiquity. This idea is sound in principle, provided that all the other natural variables remain constant. Soil permeability, rainfall, temperature, and the concentration of fluorine in groundwater all vary with time and location, however. Fluorine dating is therefore not the simple procedure that Middleton envisioned.

Still, the idea that hydroxyapatite in buried bone undergoes gradual change to fluorapatite is a correct one. In a restricted locality where there is uniformity of climate and soil, the extent of fluorine addition is at least a measure of relative age and has been so used with notable success in dating certain hominid remains. Both the Piltdown hoax, for example, and the intrusive burial of the Galley Hill skeleton were exposed in part by fluorine measurements. Supplementing them were analyses of uranium, which resembles fluorine in its increase with time, and nitrogen, which decreases as bone protein decays away.

Fluorine changes could conceivably be calibrated if bone samples were found in a radiometrically dated sequence. Conditions governing fluorine uptake, however, are so variable even over short distances that it is risky to use fluorine content as an absolute chronometer much beyond the calibration site itself. In short, fluorine dating is not now and probably never will be an absolute chronometer. Even when used in relative dating, many fluorine analyses on diverse samples are needed, and these must be supplemented by uranium and nitrogen measurements to establish confidence in the chronological conclusions.

Geomagnetic variations. Based on three centuries of direct measurement, the Earth's magnetic field is known to be varying slowly in both its intensity and direction. In fact, change seems to have been the rule throughout all of the Earth's past. Magnetic minerals in rocks (and in articles of fired clay) provide the record of ancient change, for they took on the magnetic field existing at the time of their creation or emplacement.

Polar reversals were originally discovered in lava rocks and since have been noted in deep-sea cores. In both cases the time dimension is added through radiometric methods applied to the same materials that show the reversals. Potassium–argon is the commonest chronometer used (see below *Potassium–argon methods*). A magnetic-polarity (or paleomagnetic) time scale has been proposed along the line of the geologic time scale; time divisions are called intervals, or epochs.

BIOLOGICAL PROCESSES AS ABSOLUTE CHRONOMETERS

Tree-ring growth. In the early 1900s an American astronomer named Andrew E. Douglass went looking for terrestrial records of past sunspot cycles and not only found what he sought but also discovered a useful dating method in the process. The focus of his attention was the growth rings in trees—living trees, dead trees, beams in ancient structures, and even large lumps of charcoal.

Dendrochronology

The key documents for tree-ring dating, or dendrochronology, are those trees that grow or grew where roots receive water in direct proportion to precipitation. Under such a situation, the annual tree rings vary in width as a direct reflection of the moisture supplied. What is important in tree-ring dating is the sequence in which rings vary. Suppose, for example, that a 100-year-old tree is cut down and its ring widths are measured. The results can be expressed graphically, and, if a similar graph were made from a small stump found near the 100-year-old tree, the two graphs could be compared until a match of the curves was obtained. The time when the small stump was made would thereby be determined from the position of its outer ring alongside the 100-year record.

Not every tree species nor even every specimen of a suitable species can be used. In the American Southwest, success has been achieved with yellow pine, Douglas fir, and even sagebrush. Unfortunately, the giant sequoia of California does not live in a sufficiently sensitive environment to provide a useful record. The even older bristlecone pine in California's White Mountains does have a climate-sensitive record, but its area of growth is so limited and so inaccessible that no bristlecone specimens have so far appeared in archaeological sites. This shortcoming notwithstanding, dead bristlecone pine trees are presently providing rings as old as 8,200 years for dating by carbon-14. The purpose is to check the carbon-14 method (see below *Carbon-14 dating and other cosmogenic methods*).

Coral growth. Certain fossil corals have long been used to date rocks relatively, but only recently has it been shown that corals may also serve as absolute geochronometers. They may do so by preserving a record of how many days there were in a year at the time they were growing. The number of days per year has decreased through time because the rate of rotation of the Earth has decreased; geophysical evidence suggests that days are currently lengthening at the rate of 20 seconds per million years. If this were typical of the slowdown during the past, a year consisted of 423 days about 600 million years ago.

Readings from horn corals

It is thought that horn corals indicate the number of days per year by means of their exceedingly fine external ridges of calcium carbonate, each of which is believed to represent a day's growth. Several hundred of the fine ridges also seem to cluster as a unit that presumably corresponds to one year. In certain modern West Indian corals the number of fine ridges in a presumed annual increment is approximately 360, suggesting that coral patterns are being properly interpreted.

Not many fossil corals are in a state of preservation that permits the counting of ridges, but those that are seem to lend themselves well to this procedure. Several Middle Devonian corals indicate between 385 and 410 ridges, with an average of about 400. It remains to be seen whether this method of dating, so elegant in concept and so simple in application, will blossom or wither away in the years to come. (E.A.O./Ed.)

RELATIVE AND ABSOLUTE DATING

General considerations

DISTINCTIONS BETWEEN RELATIVE-AGE AND ABSOLUTE-AGE MEASUREMENTS

Relative age

Local relationships on a single outcrop or archaeological site can often be interpreted to deduce the sequence in which the materials were assembled. This then can be used to deduce the sequence of events and processes that took place or the history of that brief period of time as recorded in the rocks or soil. For example, the presence of recycled bricks at an archaeological site indicates the sequence in which the structures were built. Similarly, in geology, if distinctive granitic pebbles can be found in the sediment beside a similar granitic body, it can be inferred that the granite, after cooling, had been uplifted and eroded and therefore was not injected into the adjacent rock sequence. Although with clever detective work many complex time sequences or relative ages can be deduced, the ability to show that objects at two separated sites were formed at the same time requires additional information. A coin, vessel, or other common artifact could link two archaeological sites, but the possibility of recycling would have to be considered. It should be emphasized that linking sites together is essential if the nature of an ancient society is to be understood, as the information at a single location may be relatively insignificant by itself. Similarly, in geologic studies, vast quantities of information from widely spaced outcrops have to be integrated. Some method of correlating rock units must be found. In the ideal case, the geologist will discover a single rock unit with a unique collection of easily observed attributes called a marker horizon that can be found at widely spaced localities. Any feature, including colour variations, textures, fossil content, mineralogy, or any unusual combinations of these can be used. It is only by correlations that the conditions on different parts of the Earth at any particular stage in its history can be deduced. In addition, because sediment deposition is not continuous and much rock material has been removed by erosion, the fossil record from many localities has to be integrated before a complete picture of the evolution of life on Earth can be assembled. Using this established record, geologists have been able to piece together events over the past 600 million years, or about one-eighth of Earth history, during which time useful fossils have been abundant. The need to correlate over the rest of geologic time, to correlate nonfossiliferous units, and to calibrate the fossil time scale has led to the development of a specialized field that makes use of natural radioactive isotopes in order to calculate absolute ages.

Absolute age

The precise measure of geologic time has proven to be the essential tool for correlating the global tectonic processes (see below) that have taken place in the past. Precise isotopic ages are called absolute ages, since they date the timing of events not relative to each other but as the time elapsed between a rock-forming event and the present. Absolute dating by means of uranium and lead isotopes has been improved to the point that for rocks 3 billion years old geologically meaningful errors of ± 1 or 2 million years can be obtained. The same margin of error applies for younger fossiliferous rocks, making absolute dating comparable in precision to that attained using fossils. To achieve this precision, geochronologists have had to develop the ability to isolate certain high-quality minerals that can be shown to have remained closed to migration of the radioactive parent atoms they contain and the daughter atoms formed by radioactive decay over billions of years of geologic time. In addition, they have had to develop special techniques with which to dissolve these highly refractory minerals without contaminating the small amount (about one-billionth of a gram) of contained lead and uranium on which the age must be calculated. Since parent uranium atoms change into daughter atoms with time at a known rate, their relative abundance leads directly to the absolute age of the host mineral. Just as the use of the fossil record has allowed a precise definition of geologic processes in approximately the past 600 million years, absolute ages allow correlations back to the Earth's oldest known rocks formed almost 4 billion years ago. In fact, even in younger rocks, absolute dating is the only way that the fossil record can be calibrated. Without absolute ages, investigators could only determine which fossil organisms lived at the same time and the relative order of their appearance in the correlated sedimentary rock record.

Unlike ages derived from fossils, which occur only in sedimentary rocks, absolute ages are obtained from minerals that grow as liquid rock bodies cool at or below the surface. When rocks are subjected to high temperatures and pressures in mountain roots formed where continents collide, certain datable minerals grow and even regrow to record the timing of such geologic events. When these regions are later exposed in uptilted portions of ancient continents, a history of terrestrial rock-forming events can be deduced. Episodes of global volcanic activity, rifting of continents, folding, and metamorphism are defined by absolute ages. The results suggest that the present-day global tectonic scheme was operative in the distant past as well.

THE GLOBAL TECTONIC ROCK CYCLE

Theory of plate tectonics

Bringing together virtually all geologic aspects of the Earth's outer rock shell (the lithosphere) into a unifying theory called plate tectonics has had a profound impact on the scientific understanding of our dynamic planet. Continents move, carried on huge slabs, or plates, of dense rock about 100 kilometres thick over a low-friction, partially melted zone (the asthenosphere) below. In the oceans, new seafloor, created at the globe-circling oceanic ridges, moves away, cools, and sinks back into the mantle in what are known as subduction zones (*i.e.,* long, narrow belts at which one plate descends beneath another). Where this occurs at the edge of a continent, as along the west coast of North and South America, large mountain chains develop with abundant volcanoes and their subvolcanic equivalents. These units, called igneous rock, or magma in their molten form, constitute major crustal additions. By contrast, crustal destruction occurs at the margins of two colliding continents, as, for example, where the subcontinent of India is moving north over Asia. Great uplift, accompanied by rapid erosion, is taking place and large sediment fans are being deposited in the Indian Ocean to the south. With time, water-soluble "cement" will cause the sandy units to become sandstone. Rocks of this kind in the ancient record may very well have resulted from rapid uplift and continent collision.

When continental plates collide, the edge of one plate is thrust onto that of the other. The rocks in the lower slab undergo changes in their mineral content in response to heat and pressure and will probably become exposed at the surface again some time later. Rocks converted to new mineral assemblages because of changing temperatures and pressures are called metamorphic. Virtually any rock now seen forming at the surface can be found in exposed deep crustal sections in a form that reveals through its mineral content the temperature and pressure of burial.

Such regions of the crust may even undergo melting and subsequent extrusion of melt magma, which may appear at the surface as volcanic rocks or may solidify as it rises to form granites at high crustal levels. Magmas produced in this way are regarded as recycled crust, whereas others extracted by partial melting of the mantle below are considered primary. (For additional information on the tectonic rock cycle, see EARTH, THE: *The surface of the Earth as a mosaic of plates: Activity along plate boundaries.*)

Even the oceans and atmosphere are involved in this great cycle because minerals formed at high temperatures are unstable at surface conditions and eventually break down or weather, in many cases taking up water and carbon dioxide to make new minerals. If such minerals were deposited on a downgoing (*i.e.,* subducted) oceanic slab, they would eventually be heated and changed back into high-temperature minerals, with their volatile components being released. These components would then rise and be fixed in the upper crust or perhaps reemerge at the surface. Such hot circulating fluids can dissolve metals and eventually deposit them as economic mineral deposits on their way to the surface.

Geochronological studies have provided documentary evidence that these rock-forming and rock-re-forming processes were active in the past. Seafloor spreading has been traced, by dating minerals found in a unique grouping of rock units thought to have been formed at the oceanic ridges, to 500 million years ago, with rare occurrences as early as 2 billion years ago. Volcanic units resembling those formed over oceanic subduction zones can be dated worldwide to show that the Earth's most prolific volcanic event occurred about 2.7 billion years ago. Other ancient volcanic units document various cycles of mountain building. The source of ancient sediment packages like those presently forming off India can be identified by dating single detrital grains of zircon found in sandstone. Magmas produced by the melting of older crust can be identified because their zircons commonly contain inherited older cores. Episodes of continental collision can be dated by isolating new zircons formed as the buried rocks underwent local melting. Periods of deformation associated with major collisions cannot be directly dated if no new minerals have formed. The time of deformation can be bracketed, however, if datable units, which both predate and postdate it, can be identified. The timing of cycles involving the expulsion of fluids from deep within the crust can be ascertained by dating new minerals formed at high pressures in exposed deep crustal sections. In some cases, it is possible to prove that gold deposits may have come from specific fluids if the deposition time of the deposits can be determined and the time of fluid expulsion is known.

Dating of volcanic units

Where the crust is under tension, as in Iceland, great fissures develop. These fissures serve as conduits that allow black lava, called basalt, to reach the surface. The portion that remains in a fissure below the surface usually forms a vertical black tubular body known as a dike (or dyke). Precise dating of such dikes can reveal times of crustal rifting in the past. Dikes and lava, now exposed on either side of Baffin Bay, have been dated to determine the time when Greenland separated from North America—namely, about 60 million years ago.

Combining knowledge of the Earth processes observed today with absolute ages of ancient geologic analogues seems to indicate that the oceans and atmosphere were present by at least 3.5 billion years ago and that they were probably released by early heating of the planet (see below *Geologic history of the Earth: Development of the atmosphere and oceans*). The continents were produced over time; the earliest portions were formed nearly 4 billion years ago, and the process still continues today. Absolute dating allows rock units formed at the same time to be identified and reassembled into ancient mountain belts, which in many cases have been disassociated by subsequent tectonic processes. The most obvious of these is the Appalachian chain that occupies the east coast of North America and extends to parts of Newfoundland as well as parts of Ireland, England, and Norway. Relic oceanic crust, formed between 480 and 500 million years ago, was identified on both sides of the Atlantic in this chain, as were numerous correlative volcanic and sedimentary units. Evidence based on geologic description, fossil content, and absolute and relative ages leave no doubt that these rocks were all part of a single mountain belt before the Atlantic Ocean opened in stages from about 200 million years ago.

DETERMINATION OF SEQUENCE

Relative geologic ages can be deduced in rock sequences consisting of sedimentary, metamorphic, or igneous rock units. In fact, they constitute an essential part in any precise isotopic, or absolute, dating program. Such is the case because most rocks simply cannot be isotopically dated. Therefore, a geologist must first determine relative ages and then locate the most favourable units for absolute dating. It is also important to note that relative ages are inherently more precise, since two or more units deposited minutes or years apart would have identical absolute ages but precisely defined relative ages. While absolute ages require expensive, complex analytical equipment, relative ages can be deduced from simple visual observations.

Inherent preciseness of relative ages

Most methods for determining relative geologic ages are well illustrated in sedimentary rocks. These rocks cover roughly 75 percent of the surface area of the continents, and unconsolidated sediments blanket most of the ocean floor. They provide evidence of former surface conditions and the life-forms that existed under those conditions. The sequence of a layered sedimentary series is easily defined because deposition always proceeds from the bottom to the top. This principle would seem self-evident, but its first enunciation more than 300 years ago by Nicolaus Steno represented an enormous advance in understanding. Known as the principle of superposition, it holds that in a series of sedimentary layers or superposed lava flows the oldest layer is at the bottom, and layers from there upward become progressively younger (see above *Study of the rock record: Early views and discoveries: The principle of superposition of rock strata*). On occasion, however, deformation may have caused the rocks of the crust to tilt, perhaps to the point of overturning them. Moreover, if erosion has blurred the record by removing substantial portions of the deformed sedimentary rock, it may not be at all clear which edge of a given layer is the original top and which is the original bottom.

Identifying top and bottom is clearly important in sequence determination, so important in fact that a considerable literature has been devoted to this question alone. Many of the criteria of top–bottom determination are based on asymmetry in depositional features. Oscillation ripple marks, for example, are produced in sediments by water sloshing back and forth. When such marks are preserved in sedimentary rocks, they define the original top and bottom by their asymmetric pattern. Certain fossils also accumulate in a distinctive pattern or position that serves to define the top side.

In wind-blown or water-lain sandstone, a form of erosion during deposition of shifting sand removes the tops of mounds to produce what are called cross-beds. The truncated layers provide an easily determined depositional top direction. The direction of the opening of mud cracks or rain prints can indicate the uppermost surface of mudstones formed in tidal areas. When a section of rock is uplifted and eroded, as during mountain-building episodes, great volumes of rock are removed, exposing a variety of differently folded and deformed rock units. The new erosion surface must postdate all units, dikes, veins, and deformation features that it crosses. Even the shapes formed on the erosional or depositional surfaces of the ancient seafloor can be used to tell which way was up. A fragment broken from one bed can only be located in a younger unit, and a pebble or animal track can only deform a preexisting unit—*i.e.,* one below. In fact, the number of ways in which one can determine the tops of well-preserved sediments is limited only by the imagination, and visual criteria can be deduced by amateurs and professionals alike.

One factor that can upset the law of superposition in major sediment packages in mountain belts is the presence of thrust faults. Such faults, which are common in compression zones along continental edges, may follow bedding planes and then cross the strata at a steep angle, placing older units on top of younger ones. In certain places, the fault planes are only a few centimetres thick and are almost impossible to detect.

Deducing relative ages in metamorphic rocks

Relative ages also can be deduced in metamorphic rocks as new minerals form at the expense of older ones in response to changing temperatures and pressures. In deep mountain roots, rocks can even flow like toothpaste in their red-hot state. Local melting may occur, and certain minerals suitable for precise isotopic dating may form both in the melt and in the host rock. In the latter case, refractory grains in particular may record the original age of the rock in their cores and the time of melting in their newly grown tips. Analytical methods are now available to date both growth stages, even though each part may weigh only a few millionths of a gram (see below *Correlation*). Rocks that flow in a plastic state record their deformation in the alignment of their constituent minerals. Such rocks then predate the deformation. If other rocks that are clearly not deformed can be found at the same site, the time of deformation can be inferred to lie between the absolute isotopic ages of the two units.

Igneous rocks provide perhaps the most striking examples of relative ages. Magma, formed by melting deep within the Earth, cuts across and hence postdates all units as it rises through the crust, perhaps even to emerge at the surface as lava. Black lava, or basalt, the most common volcanic rock on Earth, provides a simple means for determining the depositional tops of rock sequences as well as proof of the antiquity of the oceans. Pillow shapes are formed as basaltic lava is extruded (*i.e.,* erupted) under water; these are convex upward with a lower tip that projects down between two convex tops below. The shapes of pillows in ancient basalts provide both a direct indication of depositional top and proof of underwater eruption. They are widespread in rocks as old as 3.5 billion years, implying that the oceans were already present.

Basaltic lava rocks that are common where ancient continents have been rifted apart are fed from below by near vertical fractures penetrating the crust. Material that solidifies in such cracks remains behind as dikes. Here the dikes must be younger than all other units. A more interesting case develops when a cooled older crust is fractured, invaded by a swarm of dikes, and subsequently subjected to a major episode of heating with deformation and intrusion of new magma. In this instance, even though the resulting outcrop pattern is extremely complex, all of the predike units can be distinguished by the relic dikes present. The dikes also record in their newly formed minerals components that can be analyzed to give both the absolute age and the temperature and pressure of the second event. Because dike swarms are commonly widespread, the conditions determined can often be extrapolated over a broad region. Dikes do not always continue upward in a simple fashion. In some cases, they spread between the layers of near-horizontal sedimentary or volcanic units to form bodies called sills. In this situation, fragments of the host rock must be found within the intrusive body to establish its relatively younger age.

Applying the principle of crosscutting relationships

Once most or all of the relative ages of various strata have been determined in a region, it may be possible to deduce that certain units have been offset by movement along fractures or faults while others have not. Dikes that cross fault boundaries may even be found. Application of the simple principle of crosscutting relationships can allow the relative ages of all units to be deduced. A number of criteria for establishing relative ages are illustrated in the diagram shown in Figure 2.

The principles for relative age dating described above require no special equipment and can be applied by anyone on a local or regional scale. They are based on visual observations and simple logical deductions and rely on a correlation and integration of data that occurs in fragmentary form at many outcrop locations.

CORRELATION

Principles and techniques. Correlation is, as mentioned earlier, the technique of piecing together the informational content of separated outcrops. When information derived from two outcrops is integrated, the time interval they represent is probably greater than that of each alone. Presumably if all the world's outcrops were integrated, sediments representing all of geologic time would be available for examination. This optimistic hope, however, must be tempered by the realization that much of the Precambrian record—older than 570 million years—is missing. Correlating two separated outcrops means establishing that they share certain characteristics indicative of contemporary formation. The most useful indication of time equivalence is similar fossil content, provided of course that such remains are present. The basis for assuming that like fossils indicate contemporary formation is faunal succession. However, as previously noted, times of volcanism and metamorphism, which are both critical parts of global processes, cannot be correlated by fossil content. Furthermore, useful fossils are either rare or totally absent in rocks from Precambrian time, which constitutes more than 87 percent of Earth history. Precambrian rocks must therefore be correlated by means of precise isotopic dating.

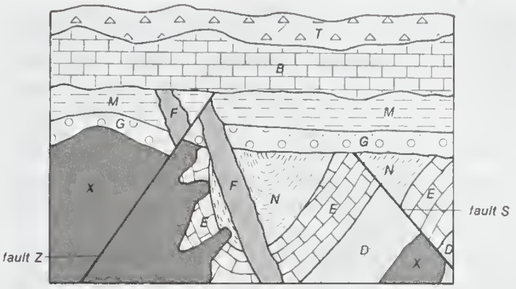


Figure 2: Hypothetical outcrop to which sequence-determining techniques can be applied. The correct sequence, from oldest to youngest, is *D,E,N,X,S,G,M,F,Z,B,T.*

From *Journal of Geological Education* (December 1967)

Unlike the principles of superposition and crosscutting, faunal succession is a secondary principle. That is to say, it depends on other sequence-determining principles for establishing its validity. Suppose there exist a number of fossil-bearing outcrops each composed of sedimentary layers that can be arranged in relative order, primarily based on superposition. Suppose, too, that all the layers contain a good representation of the animal life existing at the time of deposition. From an examination of such outcrops with special focus on the sequence of animal forms comes the empirical generalization that the faunas of the past have followed a specific order of succession, and so the relative age of a fossiliferous rock is indicated by the types of fossils it contains.

As was mentioned at the outset of this article, William Smith first noticed around 1800 that the different rock layers he encountered in his work were characterized by different fossil assemblages. Using fossils simply for identification purposes, Smith constructed a map of the various surface rocks outcropping throughout England, Wales, and southern Scotland. Smith's geologic map was extremely crude, but in its effect on Earth study it was a milestone.

Following Smith's pioneering work, generations of geologists have confirmed that similar and even more extensive fossil sequences exist elsewhere. To this day, fossils are useful as correlation tools to geologists specializing in stratigraphy. In dating the past, the primary value of fossils lies within the principle of faunal succession: each interval of geologic history had a unique fauna that associates a given fossiliferous rock with that particular interval.

Index fossils

The basic conceptual tool for correlation by fossils is the index, or guide, fossil. Ideally, an index fossil should be such as to guarantee that its presence in two separated rocks indicates their synchroneity. This requires that the lifespan of the fossil species be but a moment of time relative to the immensity of geologic history. In other words, the fossil species must have had a short temporal range. On the practical side, an index fossil should be distinctive in appearance so as to prevent misidentification, and it should be cosmopolitan both as to geography and as to rock type. In addition, its fossilized population should be sufficiently abundant for discovery to be highly probable. Such an array of attributes represents an ideal, and much stratigraphic geology is rendered difficult because of departure of the natural fossil assemblage from this ideal. Nevertheless, there is no greater testimony to the validity of fossil-based stratigraphic geology than the absolute dates made possible through radioactive measurements. Almost without exception, the relative order of strata defined by fossils has been confirmed by radiometric ages.

Correlation based on the physical features of the rock record also has been used with some success, but it is restricted to small areas that generally extend no more than several hundred kilometres. The first step is determining whether similar beds in separated outcrops can actually be traced laterally until they are seen to be part of the same original layer. Failing that, the repetition of a certain layered sequence (*e.g.,* a black shale sandwiched between a red sandstone and a white limestone) lends confidence to physical correlation. Finally, the measurement of a host of rock properties may well be the ultimate key to correlation of separated outcrops. The more ways in which two rocks are physically alike, the more likely it is that the two formed at the same time.

Only a partial listing of physical characteristics is necessary to indicate the breadth of approach in this area. Such features as colour, ripple marks, mud cracks, raindrop imprints, and slump structures are directly observable in the field. Properties derived from laboratory study include (1) size, shape, surface appearance, and degree of sorting of mineral grains, (2) specific mineral types present and their abundances, (3) elemental composition of the rock as a whole and of individual mineral components, (4) type and abundance of cementing agent, and (5) density, radioactivity, and electrical-magnetic-optical properties of the rock as a whole.

Well logging

With the development of miniaturized analytical equipment, evaluation of rock properties down a small drill hole has become possible. The technique, called well logging, involves lowering a small instrument down a drill hole on the end of a wire and making measurements continuously as the wire is played out in measured lengths. By this technique it is possible to detect depth variations in electrical resistivity, self-potential, and gamma-ray emission rate and to interpret such data in terms of continuity of the layering between holes. Subsurface structures can thus be defined by the correlation of such properties.

Field geologists always prize a layer that is so distinctive in appearance that a series of tests need not be made to establish its identity. Such a layer is called a key bed. In a large number of cases, key beds originated as volcanic ash. Besides being distinctive, a volcanic-ash layer has four other advantages for purposes of correlation: it was laid down in an instant of geologic time; it settles out over tremendous areas; it permits physical correlation between contrasting sedimentary environments; and unaltered mineral crystals that permit radiometric measurements of absolute age often are present.

Correlation may be difficult or erroneous if several different ash eruptions occurred, and a layer deposited in one is correlated with that from another. Even then, the correlation may be justified if the two ash deposits represent the same volcanic episode. Much work has been undertaken to characterize ash layers both physically and chemically and so avoid incorrect correlations. Moreover, single or multigrain zircon fractions from the volcanic source are now being analyzed to provide precise absolute ages for the volcanic ash and the fossils in the adjacent units.

Geologic column and its associated time scale. The end product of correlation is a mental abstraction called the geologic column. It is the result of integrating all the world's individual rock sequences into a single sequence. In order to communicate the fine structure of this so-called column, it has been subdivided into smaller units. Lines are drawn on the basis of either significant changes in fossil forms or discontinuities in the rock record (*i.e.,* unconformities, or large gaps in the sedimentary sequence); the basic subdivisions of rock are called systems, and the corresponding time intervals are termed periods. In the upper part of the geologic column, where fossils abound, these rock systems and geologic periods are the basic units of rock and time. Lumping of periods results in eras, and splitting gives rise to epochs. In both cases, a threefold division into early–middle–late is often used, although those specific words are not always applied. Similarly, many periods are split into three epochs. Names assigned to individual epochs follow no single worldwide standard except for the seven epochs making up the last two periods (see below Table 4).

Over the interval from the Paleozoic to the present, about 35 epochs are recognized in North America. This interval is represented by approximately 250 formations, discrete layers thick enough and distinctive enough in lithology to merit delineation as units of the geologic column. Also employed in subdivision is the zone concept, in which it is the fossils in the rocks rather than the lithologic character that defines minor stratigraphic boundaries. The basis of zone definition varies among geologists, some considering a zone to be all rocks containing a certain species (usually an invertebrate), whereas others focus on special fossil assemblages.

The lower part of the geologic column, where fossils are very scarce, was at one time viewed in the context of two eras of time, but subsequent mapping has shown the provincial bias in such a scheme. Consequently, the entire lower column is now considered a single unit, the Precambrian. The results of isotopic dating are now providing finer Precambrian subdivisions that have worldwide applicability. For a detailed discussion of this subject, see below *Precambrian era.*

The geologic column and the relative geologic time scale are sufficiently defined to fulfill the use originally envisioned for them—providing a framework within which to tell the story of Earth history. Just as human history has its interweaving plots of warfare, cultural development, and technological advance, so the Earth's rocks tell another story of intertwined sequences of events. Mountains have been built and eroded away, seas have advanced and retreated, a myriad of life-forms has inhabited land and sea. In all these happenings the geologic column and its associated time scale spell the difference between an unordered series of isolated events and the unfolding story of a changing Earth.

Absolute dating

Although relative ages can generally be established on a local scale, the events recorded in rocks from different locations can be integrated into a picture of regional or global scale only if their sequence in time is firmly established. The time that has elapsed since certain minerals formed can now be determined because of the presence of a small amount of natural radioactive atoms in their structures. Whereas studies using fossil dating began almost 300 years ago, radioactivity itself was not discovered until roughly a century ago, and it has only been from about 1950 that extensive efforts to date geologic materials have become common. Methods of isotopic measurement continue to be refined today, and absolute dating has become an essential component of virtually all field-oriented geologic investigations. In the process of refining isotopic measurements, methods for low-contamination chemistry had to be developed, and it is significant that many such methods now in worldwide use resulted directly from work in geochronology.

It has already been explained how different Earth processes create different rocks as part of what can be considered a giant rock-forming and -reforming cycle. Attention has been called wherever possible to those rocks that contain minerals suitable for precise isotopic dating. It is important to remember that precise ages cannot be obtained for just any rock unit but that any unit can be dated relative to a datable unit. The following discussion will show why this is so, treating in some detail the analytic and geologic problems that have to be overcome if precise ages are to be determined. It will become apparent, for example, that isotopic ages can be reset by high temperatures; however, this seeming disadvantage can be turned to one's favour in determining the cooling history of a rock. As various dating methods are discussed, the great interdependence of the geologic and analytic components essential to geochronology should become evident.

Significance of isotope geology

The field of isotope geology complements geochronology. Workers in isotope geology follow the migration of isotopes produced by radioactive decay through large- and small-scale geologic processes. Isotopic tracers of this kind can be thought of as an invisible dye injected by nature into Earth systems that can be observed only with sophisticated instruments. Studying the movement or distribution of these isotopes can provide insights into the nature of geologic processes.

PRINCIPLES OF ISOTOPIC DATING

All absolute isotopic ages are based on radioactive decay, a process whereby a specific atom or isotope is converted into another specific atom or isotope at a constant and known rate. Most elements exist in different atomic forms that are identical in their chemical properties but differ in the number of neutral particles—*i.e.*, neutrons—in the nucleus. For a single element, these atoms are called isotopes. Because isotopes differ in mass, their relative abundance can be determined if the masses are separated in a mass spectrometer (see below *Use of mass spectrometers*).

Observing the radioactive decay process in the laboratory

Radioactive decay can be observed in the laboratory by either of two means: (1) a radiation counter (*e.g.*, a Geiger counter), which detects the number of high-energy particles emitted by the disintegration of radioactive atoms in a sample of geologic material, or (2) a mass spectrometer, which permits the identification of daughter atoms formed by the decay process in a sample containing radioactive parent atoms. The particles given off during the decay process are part of a profound fundamental change in the nucleus. To compensate for the loss of mass (and energy), the radioactive atom undergoes internal transformation and in most cases simply becomes an atom of a different chemical element. In terms of the numbers of atoms present, it is as if apples changed spontaneously into oranges at a fixed and known rate. In this analogy, the apples would represent radioactive, or parent, atoms, while the oranges would represent the atoms formed, the so-called daughters. Pursuing this analogy further, one would expect that a new basket of apples would have no oranges but that an older one would have many. In fact, one would expect that the ratio of oranges to apples would change in a very specific way over the time elapsed, since the process continues until all the apples are converted. In geochronology the situation is identical. A particular rock or mineral that contains a radioactive isotope (or radioisotope) is analyzed to determine the number of parent and daughter isotopes present, whereby the time since that mineral or rock formed is calculated. Of course, one must select geologic materials that contain elements with long half-lives—*i.e.*, those for which some parent atoms would remain.

Given below is the simple mathematical relationship that allows the time elapsed to be calculated from the measured parent/daughter ratio. The age calculated is only as good as the existing knowledge of the decay rate and is valid only if this rate is constant over the time that elapsed.

Radioactive decay as an immutable process

Fortunately for geochronology the study of radioactivity has been the subject of extensive theoretical and laboratory investigation by physicists for almost a century. The results show that there is no known process that can alter the rate of radioactive decay. By way of explanation it can be noted that since the cause of the process lies deep within the atomic nucleus, external forces such as extreme heat and pressure have no effect. The same is true regarding gravitational, magnetic, and electric fields, as well as the chemical state in which the atom resides. In short, the process of radioactive decay is immutable under all known conditions. Although it is impossible to predict when a particular atom will change, given a sufficient number of atoms, the rate of their decay is found to be constant. The situation is analogous to the death rate among human populations insured by an insurance company. Even though it is impossible to predict when a given policyholder will die, the company can count on paying off a certain number of beneficiaries every month. The recognition that the rate of decay of any radioactive parent atom is proportional to the number of atoms (N) of the parent remaining at any time gives rise to the following expression:

$$\begin{array}{ccc} R & \propto & N \\ \text{rate of} & \text{is propor-} & \text{number of parent} \\ \text{disintegration} & \text{tional to} & \text{atoms present.} \end{array} \qquad (1)$$

Decay constant

Converting this proportion to an equation incorporates the additional observation that different radioisotopes have different disintegration rates even when the same number of atoms are observed undergoing decay. In other words, each radioisotope has its own decay constant, abbreviated λ, which provides a measure of its intrinsic rapidity of decay. Proportion 1 becomes:

$$R = \lambda N. \qquad (2)$$

Stated in words, this equation says that the rate at which a certain radioisotope disintegrates depends not only on how many atoms of that isotope are present but also on an intrinsic property of that isotope represented by λ, the so-called decay constant. Values of λ vary widely—from 10^{20} reciprocal seconds (*i.e.*, the unit of 1 second) for a rapidly disintegrating isotope such as helium-5 to less than 10^{-25} reciprocal seconds for slowly decaying cerium-142.

In the calculus, the rate of decay R in equation 2 is written as the derivative dN/dt, in which dN represents the small number of atoms that decay in an infinitesimally short time interval dt. Replacing R by its equivalent dN/dt results in the differential equation

$$\frac{dN}{dt} = -\lambda N. \qquad (3)$$

Solution of this equation by techniques of the calculus yields one form of the fundamental equation for radiometric age determination,

$$\frac{N}{N_0} = e^{-\lambda t}, \qquad (4)$$

in which N_0 is the number of radioactive atoms present in a sample at time zero, N is the number of radioactive atoms present in the sample today, e is the base of natural logarithms (equal to about 2.72), λ is the decay constant of the radioisotope being considered, and t is the time elapsed since time zero.

Definition of half-life

Two alterations are generally made to equation 4 in order to obtain the form most useful for radiometric dating. In the first place, since the unknown term in radiometric dating is obviously t, it is desirable to rearrange equation 4 so that it is explicitly solved for t. Second, the more common way to express the intrinsic decay rate of a radioisotope is through its half-life (abbreviated $t_{1/2}$) rather than through the decay constant λ. Half-life is defined as the time period that must elapse in order to halve the initial number of radioactive atoms. The half-life and the decay constant are inversely proportional because rapidly decaying radioisotopes have a high decay constant but a short half-life. With t made explicit and half-life introduced, equation 4 is converted to the following form, in which the symbols have the same meaning:

$$t = \frac{t_{1/2}}{0.693} \times \log_e\left(\frac{N_0}{N}\right). \qquad (5)$$

Alternatively, because the number of daughter atoms is directly observed rather than N, which is the initial number of parent atoms present, another formulation may be more convenient. Since the initial number of parent atoms present at time zero N_0 must be the sum of the parent atoms remaining N and the daughter atoms present D, one can write:

$$D = N_0 - N. \qquad (6)$$

From equation 4 above, it follows that $N_0 = N(e^{\lambda t})$. Substituting this in equation 6 gives

$$D = Ne^{\lambda t} - N,$$

or

$$D = N(e^{\lambda t} - 1),$$

or $$\frac{D}{N} = (e^{\lambda t} - 1).$$

If one chooses to use P to designate the parent atom, the expression assumes its familiar form:

$$\frac{D}{P} = (e^{\lambda t} - 1) \tag{7}$$

and $$t = \frac{1}{\lambda} \ln \left(\frac{D}{P} + 1 \right). \tag{8}$$

This pair of equations states rigorously what might be assumed from intuition, that minerals formed at successively longer times in the past would have progressively higher daughter-to-parent ratios. This follows because, as each parent atom loses its identity with time, it reappears as a daughter atom. The increase in D/P with time is evident in equation (7) because larger values of time will increase the value of $e^{\lambda t}$, where λ is constant. Equation (8) documents the simplicity of direct isotopic dating. The time of decay is proportional to the natural logarithm (represented by ln) of the ratio of D to P. In short, one need only measure the ratio of the number of radioactive parent and daughter atoms present, and the time elapsed since the mineral or rock formed can be calculated, provided of course that the decay rate is known. Likewise, the conditions that must be met to make the calculated age precise and meaningful are in themselves simple:

Conditions for precise age calculations

1. The rock or mineral must have remained closed to the addition or escape of parent and daughter atoms since the time that the rock or mineral (system) formed.
2. It must be possible to correct for other atoms identical to daughter atoms already present when the rock or mineral formed.
3. The decay constant must be known.
4. The measurement of the daughter-to-parent ratio must be accurate because uncertainty in this ratio contributes directly to uncertainty in the age.

Different schemes have been developed to deal with the critical assumptions stated above. In uranium–lead dating, minerals virtually free of initial lead can be isolated and corrections made for the trivial amounts present. In whole rock isochron methods that make use of the rubidium–strontium or samarium–neodymium decay schemes (see below), a series of rocks or minerals are chosen that can be assumed to have the same age and identical abundances of their initial isotopic ratios. The results are then tested for the internal consistency that can validate the assumptions. In all cases, it is the obligation of the investigator making the determinations to include enough tests to indicate that the absolute age quoted is valid within the limits stated. In other words, it is the obligation of geochronologists to try to prove themselves wrong by including a series of cross-checks in their measurements before they publish a result. Such checks include dating a series of ancient units with closely spaced but known relative ages, as shown in Figure 3, and replicate analysis of different parts of the same rock body with samples collected at widely spaced localities.

The importance of internal checks as well as interlaboratory comparisons becomes all the more apparent when one realizes that geochronology laboratories are limited in number. Because of the expensive equipment necessary and the combination of geologic, chemical, and laboratory skills required, geochronology is usually carried out by teams of experts. Most geologists must rely on geochronologists for their results. In turn, the geochronologist relies on the geologist for relative ages. (E.A.O./T.E.Kr.)

From D W Davis and J B Paces, *Eerth end Planetary Science Letters*, vol 97, no. 1/2 (February 1990), Elsevier Science Publishers B V

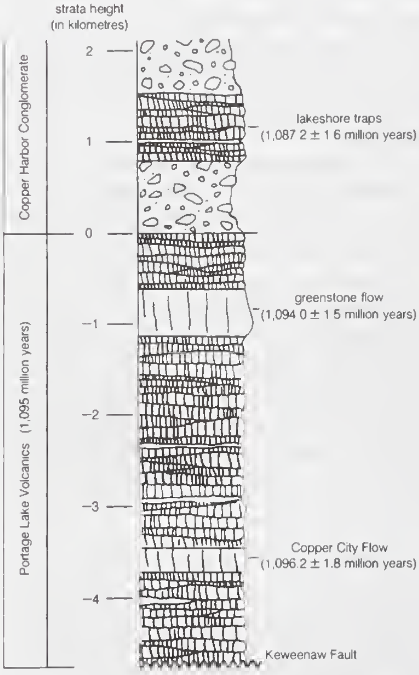

strata height
(in kilometres)
Copper Harbor Conglomerate
Portage Lake Volcanics (1,095 million years)
2
1
0
−1
−2
−3
−4
lakeshore traps
(1,087 2 ± 1 6 million years)
greenstone flow
(1,094 0 ± 1 5 million years)
Copper City Flow
(1,096.2 ± 1.8 million years)
Keweenaw Fault

Figure 3: *Absolute versus relative age, Keweenawan Lava.*
Precise absolute uranium–lead zircon ages are compared to relative ages for a series of basaltic and andesite lava flows in the Keweenawan Peninsula in Michigan. Although determined only at one location, the ages have regional significance as they bracket a widely correlated time during which the Earth's magnetic field had a reversed polarity. In addition, the ages establish the time of transition between lava eruption and conglomerate deposition (see text).

EVALUATION AND PRESENTATION SCHEMES IN DATING

Origin of radioactive elements used. In order for a radioactive parent–daughter pair to be useful for dating, many criteria must be met. This section examines these criteria and explores the ways in which the reliability of the ages measured can be assessed. Because geologic materials are diverse in their origin and chemical content and datable elements are unequally distributed, each method has its strengths and weaknesses.

When the elements in the Earth were first created, many radioactive isotopes were present. Of these, only the radioisotopes with extremely long half-lives remain. Table 2 lists a number of such isotopes and their respective daughter products that are used in various forms of rock dating. It should be mentioned in passing that some of the radioisotopes present early in the history of the solar system and now completely extinct have been recorded in meteorites in the form of the elevated abundances of their daughter isotopes. Analysis of such meteorites makes it possible to estimate the time that elapsed between element creation and meteorite formation. Natural elements that are still radioactive today produce daughter products at a very slow rate; hence, it is easy to date very old minerals but difficult to obtain the age of those formed in the recent geologic past. This follows from the fact that the amount of daughter isotopes present is so small that it is difficult to measure. The difficulty can be overcome to some degree by achieving lower background contamination, by improving instrument sensitivity, and by finding minerals with abundant parent isotopes. Geologic events of the not-too-distant past are more easily dated by using recently formed radioisotopes with short half-lives that produce more daughter products per unit time. Two sources of such isotopes exist. In one case, intermediate isotopes in the uranium or thorium decay chain can become isolated in certain minerals due to differences in chemical properties and, once fixed, can decay to new isotopes, providing a measure of the time elapsed since they were isolated. To understand this, one needs to know that though uranium-238 (^{238}U) does indeed decay to lead-206 (^{206}Pb), as indicated in Table 2, it is not a one-step process. In fact, this is a multistep process involving the expulsion of eight alpha particles and six beta particles, along with a considerable amount of energy. There exists a series of different elements, each of them in a steady state where they form at the same rate as they disintegrate. The number present is proportional to their decay rate, with long-lived members being more abundant. Because all of these isotopes have relatively short half-lives, none remains since the creation of the elements, but instead they are continuously pro-

Radioisotopes suitable for dating recent geologic events

Table 2: Major Decay Schemes for Isotopic Dating

parent isotope	daughter isotope	half-life in years	applicable materials
^{238}U	^{206}Pb	4.468 × 10^9	igneous and metamorphic rocks with zircon, baddeleyite, perovskite, monazite, titanite, rutile, xenotime, pitchblende, thorite, and thorianite; whole rock carbonates; single-mineral grains from sediments
^{235}U	^{207}Pb	0.7038 × 10^9	
^{232}Th	^{208}Pb	14.01 × 10^9	
^{40}K	^{40}Ar	1.25 × 10^9	potassium-bearing minerals (*e.g.*, mica); hornblende; meteorite impact glass; authigenic minerals
^{147}Sm	^{143}Nd	1.06 × 10^{10}	mafic igneous rocks; meteorites; metamorphic garnets
^{87}Rb	^{87}Sr	4.88 × 10^{10}	potassium-bearing minerals; authigenic minerals in sediments; felsic whole rocks
^{187}Re	^{187}Os	4.56 × 10^{10}	trace minerals from mineral deposits; molybdenite; others under investigation

vided by the decay of the long-lived parent. This type of dating, known as disequilibrium dating, will be explored below in the section *Uranium-series disequilibrium dating.*

Another special type of dating employs recently formed radioisotopes produced by cosmic-ray bombardment of target atoms at the Earth's surface or in the atmosphere. The amounts produced, although small, provide insight into many near-surface processes in the geologic past. This aspect of geology is becoming increasingly important as researchers try to read the global changes that took place during the Earth's recent past in an effort to understand or predict the future. The most widely used radioactive cosmogenic isotope is carbon of mass 14 (^{14}C), which provides a method of dating events that have occurred over roughly the past 50,000 years. This time spans much of the historic and prehistoric record of mankind. Cosmogenic isotopes and those used for disequilibrium dating are listed in Table 3.

The isochron method. Many radioactive dating methods are based on minute additions of daughter products to a rock or mineral in which a considerable amount of daughter-type isotopes already exists. These isotopes did not come from radioactive decay in the system but rather formed during the original creation of the elements. In this case, it is a big advantage to present the data in a form in which the abundance of both the parent and daughter isotopes are given with respect to the abundance of the initial background daughter. The incremental additions of the daughter type can then be viewed in proportion to the abundance of parent atoms. In mathematical terms this is achieved as follows. It has already been shown—equation 7—that the number of daughter atoms present from radioactive decay D^* can be related to the number of parent atoms remaining P by the simple expression:

$$D^* = P(e^{\lambda t} - 1). \qquad (9)$$

Table 3: Principal Cosmogenic and Uranium–Thorium Series Radioisotopes

radioisotope	half-life in years	principal uses
Cosmogenic isotope		
^{10}Be	1.5 × 10^6	dating marine sediment, manganese nodules, glacial ice, quartz in rock exposures, terrestrial age of meteorites, and petrogenesis of island-arc volcanics
^{14}C	5,730 ± 40	dating of biogenic carbon, calcium carbonate, terrestrial age of meteorites
^{26}Al	0.716 × 10^6	dating marine sediment, manganese nodules, glacial ice, quartz in rock exposures, terrestrial age of meteorites
^{32}Si	276 ± 32	dating biogenic silica, glacial ice
^{36}Cl	0.308 × 10^6	dating glacial ice, exposures of volcanic rocks, groundwater, terrestrial age of meteorites
^{39}Ar	269	dating glacial ice, groundwater
^{53}Mn	3.7 × 10^6	terrestrial age of meteorites, abundance of extraterrestrial dust in ice and sediment
^{59}Ni	8 × 10^4	terrestrial age of meteorites, abundance of extraterrestrial dust in ice and sediment
^{81}Kr	0.213 × 10^6	dating glacial ice, cosmic-ray exposure age of meteorites
U–Th series isotope		
^{234}U	2.48 × 10^5	dating coral and carbonate deposits in oceans and lakes
^{230}Th	7.52 × 10^4	dating ocean sediments
^{210}Pb	22.26	dating glacial ice and recent sediments
^{231}Pa	3.248 × 10^4	dating recent sediments

Source: Adapted from Gunter Faure, *Principles of Isotope Geology.* Copyright © 1986 by John Wiley & Sons.

When some daughter atoms are initially present (designated D_0), the total number D is the sum of radiogenic and initial atoms, so that

$$D = D_0 + P(e^{\lambda t} - 1). \qquad (10)$$

To establish the condition that both parent and daughter abundances should be relative to the initial background, a stable isotope S of the daughter element can be chosen and divided into all portions of this equation; thus,

$$\frac{D}{S} = \left(\frac{D}{S}\right)_0 + \frac{P}{S}(e^{\lambda t} - 1).$$

This equation has the form; $y = b + xm$, which is that of a straight line on x–y coordinates. The slope m is equal to $(e^{\lambda t} - 1)$ and the intercept is equal to $(D/S)_0$. This term, shown in Figure 4, is called the initial ratio. The slope is proportional to the geologic age of the system.

Advantages of the isochron method

In practice, the isochron approach has many inherent advantages. When a single body of liquid rock crystallizes, parent and daughter elements may separate so that, once solid, the isotopic data would define a series of points, such as those shown as open circles designated R$_1$, R$_2$, R$_3$ in Figure 4. They plot along a horizontal line reflecting a common value for the initial daughter isotope ratio $(D/S)_0$. With time each would then develop additional daughter abundances in proportion to the amount of parent present. If a number of samples are analyzed and the results are shown to define a straight line within error, then a precise age is defined because this is only possible if each is a closed system and each has the same initial ratio and age. The uncertainty in determining the slope is reduced because it is defined by many points. A second advantage of the method relates to the fact that under high-temperature conditions the daughter isotopes may escape from the host minerals. In this case, a valid age can still be obtained, provided that they remain within the rock. Should a point plot below the line, it could indicate that a particular sample was open to migration of the dating elements or that the sample was contaminated and lay below the isochron when the rock solidified.

Rubidium–strontium (Rb–Sr) dating was the first technique in which the whole rock isochron method was extensively employed. Certain rocks that cooled quickly at the surface were found to give precisely defined linear isochrons, but many others did not. Some studies have shown that rubidium is very mobile both in fluids that migrate through the rock as it cools and in fluids that are present as the rock undergoes chemical weathering. Similar studies have shown that the samarium–neodymium (Sm–Nd) parent–daughter pair is more resistant to secondary migration but that, in this instance, sufficient initial spread in the abundance of the parent isotope is difficult to achieve.

Analysis of separated minerals. When an igneous rock crystallizes, a wide variety of major and trace minerals may form, each concentrating certain elements and radioactive trace elements within the rock. By careful selection, certain minerals that contain little or no daughter element but abundant parent element can be analyzed. In this case, the slope of the line in Figure 4 is computed from an assumed value for the initial ratio, and it is usually possible to show that uncertainties related to this assumption are negligible. This is possible in potassium–argon (K–Ar) dating, for example, because most minerals do not take argon into their structures initially. In rubidium–strontium dating, micas exclude strontium when they form, but accept much rubidium. In uranium–lead (U–Pb) dating of zircon, the zircon is found to exclude initial lead almost completely. Minerals, too, are predictable chemical compounds that can be shown to form at specific temperatures and remain closed up to certain temperatures if a rock has been reheated or altered. A rock, on the other hand, may contain minerals formed at more than one time under a variety of conditions. Under such circumstances the isolation and analysis of certain minerals can indicate at what time these conditions prevailed. If a simple mineral is widespread in the geologic record, it is more valuable for dating as more units can be measured for age and compared by the same method. However, if a single parent–daughter pair that is amenable to precise analysis can be measured in a variety of minerals, the ages of a wide variety of rock types can be determined by a single method without the need for intercalibration. In some cases the discovery of a rare trace mineral results in a major breakthrough as it allows precise ages to be determined in formerly undatable units. For example, the mineral baddeleyite, an oxide of zirconium (ZrO_2), has been shown to be widespread in small amounts in mafic igneous rocks (*i.e.,* those composed primarily of one or more ferromagnesian, dark-coloured minerals). Here, a single uranium–lead isotopic analysis can provide an age that is more precise than can be obtained by the whole rock isochron method involving many analyses. When single minerals are analyzed, each grain can be studied under a microscope under intense side light so that alterations or imperfections can be revealed and excluded. If minerals are used for dating, the necessary checks on the ages are achieved by analyzing samples from more than one location and by analyzing different grain sizes or mineral types that respond differently to disturbing events. In summary, it can be said that minerals provide a high degree of sample integrity that can be predicted on the basis of experience gained through numerous investigations under a variety of geologic conditions. An ideal mineral is one that has sufficient parent and daughter isotopes to measure precisely, is chemically inert, contains little or no significant initial daughter isotopes, and retains daughter products at the highest possible temperatures. A specific datable mineral like rutile, which can be linked to a specific event such as the formation of a mineral deposit, is especially important.

Use of rare trace minerals for dating

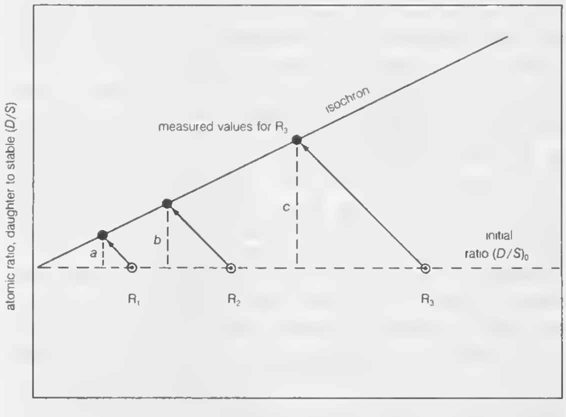

atomic ratio, daughter to stable (D/S)
measured values for R₃
isochron
a
b
c
initial ratio (D/S)₀
R₁
R₂
R₃
atomic ratio, parent to stable isotopes = P/S

Figure 4: *Isochron diagram.*
The isochron diagram allows radiogenic additions to be viewed in proportion to the abundance of radioactive parent atoms, all relative to a common background. In this case, the value of the initial parent-to-stable ratio in R_2 has been chosen to be twice that in R_1, and R_3 is set at twice that of R_2. Under these conditions, the amount of daughter isotope added over geologic time is proportional, so that the length of lines *a*, *b*, c are in the ratio of 1 to 2 to 4. Because each parent atom that disintegrates becomes a daughter atom, the point moves over and up on a −45° slope as drawn. With time, the line called an isochron rotates, and data plot at the positions shown by closed circles. These would be the values measured today, and the slope of the line would indicate the time elapsed.

Model ages. Since the Earth was formed, the abundance of daughter product isotopes, such as those listed in Table 2, has increased through time. For example, the ratio of lead of mass 206 relative to that of mass 204 has changed from an initial value of about 10 present when the Earth was formed to an average value of about 19 in rocks at the terrestrial surface today. This is true because uranium is continuously creating more lead. A lead-rich mineral formed and isolated early in Earth history would have a low lead-206 to lead-204 ratio because it did not receive subsequent additions by the radioactive decay of uranium. If the Earth's interior were a simple and homogeneous reservoir with respect to the ratio of uranium to lead, a single sample extracted by a volcano would provide the time of extraction. This would be called a model age. No parent–daughter value for a closed system is involved, rather just a single isotopic measurement of lead viewed with respect to the expected evolution of lead in the Earth. Unfortunately the simplifying assumption in this case is not true, and lead model ages are approximate at best. Other model ages can be calculated using neodymium isotopes by extrapolating present values back to a proposed mantle-evolution line. In both cases, approximate ages that have a degree of validity with respect to one another result, but they are progressively less reliable as the assumptions on which the model is calculated are violated.

The progressive increase in the abundance of daughter isotopes over time gains a special significance where the parent element is preferentially enriched in either the mantle or the crust. For example, rubidium is concentrated in the crust, and as a result the present-day continents, subjected to weathering, have an elevated radiogenic to stable isotope ratio ($^{87}Sr/^{86}Sr$) of 0.720. In contrast, modern volcanic rocks in the oceans imply that much of the mantle has a value between about 0.703 and 0.705. Should crustal material be recycled, the strontium isotopic signature of the melt would be diagnostic.

Multiple ages for a single rock; the thermal effect. Fossils record the initial, or primary, age of a rock unit. Isotopic systems, on the other hand, can yield either the primary age or the time of a later event, because crystalline materials are very specific in the types of atoms they incorporate, in terms of both the atomic size and charge. An element formed by radioactive decay is quite different from its parent atom and thus is out of place with respect to the host mineral. All it takes for such an element to be purged from the mineral is sufficient heat to allow solid diffusion to occur. Each mineral has a temperature at which rapid diffusion sets in, so that, as a region is slowly heated, first one mineral and then another loses its daughter isotopes. When this happens, the isotopic "clock" is reset to zero, where it remains until the mineral cools below the blocking temperature. (This is the temperature below which a mineral becomes a closed chemical system for a specific radioactive decay series. Accordingly, the parent–daughter isotope ratio indicates the time elapsed since that critical threshold was reached.) In this case, the host mineral could have an absolute age very much older than is recorded in the isotopic record. The isotopic age then is called a cooling age. It is even possible by using a series of minerals with different blocking temperatures to establish a cooling history of a rock body—*i.e.,* the times since the rock body cooled below successively lower temperatures. Such attempts can be complicated by the fact that a mineral may "grow" below the blocking temperature rather than simply become closed to isotopic migration. When this happens, the age has little to do with the cooling time. Another problem arises if a region undergoes a second reheating event. Certain minerals may record the first event, whereas others may record the second, and any suggestion of progressive cooling between the two is invalid. This complication does not arise when rapid cooling has occurred. Identical ages for a variety of minerals with widely different blocking temperatures is unequivocal proof of rapid cooling.

Blocking temperature

Fortunately for geologists the rock itself records in its texture and mineral content the conditions of its formation. A rock formed at the surface with no indication of deep burial or new mineral growth can be expected to give a valid primary age by virtue of minerals with low blocking temperatures. On the other hand, low-blocking-point minerals from a rock containing minerals indicative of high temperatures and pressures cannot give a valid primary age. Such minerals would be expected to remain open until deep-level rocks of this sort were uplifted and cooled.

Given these complicating factors, one can readily understand why geochronologists spend a great deal of their time and effort trying to see through thermal events that occurred after a rock formed. The importance of identifying and analyzing minerals with high blocking temperatures also cannot be overstated. Minerals with high blocking temperatures that form only at high temperatures are especially valuable. Once formed, these minerals can resist daughter loss and record the primary age even though they remained hot (say, 700° C) for a long time. The mineral zircon datable by the uranium–lead method is one such mineral. The mica mineral biotite dated by either the potassium–argon or the rubidium–strontium method occupies the opposite end of the spectrum and does not retain daughter products until cooled below about 300° C. Successively higher blocking temperatures are recorded for another mica type known as muscovite and for amphibole, but the ages of both of these minerals can be completely reset at temperatures that have little or no effect on zircon.

Taken in perspective, it is evident that many parts of the Earth's crust have experienced reheating temperatures above 300° C—*i.e.*, reset mica ages are very common in rocks formed at deep crustal levels. Vast areas within the Precambrian shield, which have identical ages reflecting a common cooling history, have been identified (see below *Precambrian rocks*). These are called geologic provinces. By contrast, rocks that have approached their melting point, say, 750° C, which can cause new zircon growth during a second thermal event, are rare, and those that have done this more than once are almost nonexistent.

INSTRUMENTS AND PROCEDURES

Use of mass spectrometers. The age of a geologic sample is measured on as little as a billionth of a gram of daughter isotopes. Moreover, all the isotopes of a given chemical element are nearly identical except for a very small difference in mass. Such conditions necessitate instrumentation of high precision and sensitivity. Both these requirements are met by the modern mass spectrometer. A high-resolution mass spectrometer of the type used today was first described by the American physicist Alfred O. Nier in 1940, but it was not until about 1950 that such instruments became available for geochronological research.

For isotopic dating with a mass spectrometer, a beam of charged atoms, or ions, of a single element from the sample is produced. This beam is passed through a strong magnetic field in a vacuum, where it is separated into a number of beams, each containing atoms of only the same mass. Because of the unit electric charge on every atom, the number of atoms in each beam can be evaluated by collecting individual beams sequentially in a device called a Faraday cup. Once in this collector, the current carried by the atoms is measured as it leaks across a resistor to ground. Currents measured are small, only from 10^{-11} to 10^{-15} ampere, so that shielding and preamplification are required as close to the Faraday cup as possible. It is not possible simply to count the atoms, because all atoms loaded into the source do not form ions and some ions are lost in transmission down the flight tube. Precise and accurate information as to the number of atoms in the sample can, however, be obtained by measuring the ratio of the number of atoms in the various separated beams. By adding a special artificially enriched isotope during sample dissolution and by measuring the ratio of natural to enriched isotopes in adjacent beams, the number of daughter isotopes can be readily determined. The artificially enriched isotope is called a "spike." It is usually a highly purified form of a low-abundance natural isotope, but an even better spike is an isotope with a mass not found in nature at all. Lead-205 produced in a type of particle accelerator called a cyclotron constitutes such an ideal spike.

Adding a spike

As the sample is heated and vaporizes under the vacuum in the source area of the mass spectrometer, it is commonly observed that the lighter isotopes come off first, causing a bias in the measured values that changes during the analysis. In most cases this bias, or fractionation, can be corrected if the precise ratio of two of the stable isotopes present is known. Today's state-of-the-art instruments produce values for strontium and neodymium isotopic abundances that are reproducible at a level of about 1 in 20,000. Such precision is often essential in the isochron method (see above) because of the small changes in relative daughter abundance that occur over geologic time.

Isotope dilution

Technical advances. The ability to add a single artificial mass to the spectrum in a known amount and to determine the abundances of other isotopes with respect to this provides a powerful analytical tool. By means of this process known as isotope dilution, invisibly small amounts of material can be analyzed, and because only ratios are involved, a loss of part of the sample during preparation has no effect on the result. Spike solutions can be calibrated simply by obtaining a highly purified form of the element being calibrated. After carefully removing surface contamination, a precisely weighted portion of the element is dissolved in highly purified acid and diluted to the desired level in a weighed quantity of water. What is required is dilution of one cubic centimetre to a litre from which a second cubic centimetre is again diluted to a litre to approach the range of parts per million or parts per billion typically encountered in samples. In this way, a known number of natural isotopes can be mixed with a known amount of spike and the concentration in the spike solution determined from the ratio of the masses. Once the calibration has been completed, the process is reversed and a weighed amount of spike is mixed with the parent and daughter elements from a mineral or rock. The ratio of the masses then gives the number of naturally produced atoms in the sample. The use of calibrated enriched isotopic tracers facilitates checks for contamination, even though the process is time-consuming. A small but known amount of tracer added to a beaker of water can be evaporated under clean-room conditions. Once loaded in a mass spectrometer, the contamination from the beaker and the water is easily assessed with respect to the amount of spike added. Contamination as small as 10^{-12} gram can be detected by this method.

The materials analyzed during isotopic investigations vary from microgram quantities of highly purified mineral grains to gram-sized quantities of rock powders. In all cases, the material must be dissolved without significant contamination. The spike should be added before dissolution. Most of the minerals in rocks can be dissolved in a day or so at a temperature near 100° C. Certain minerals that are highly refractory both in nature and in the laboratory (*e.g.*, zircon) may require five days or more at temperatures near 220° C. In this case, the sample is confined in a solid Teflon (trade name for a synthetic resin composed of polytetrafluoroethylene), metal-clad pressure vessel, introduced by the Canadian geochronologist Thomas E. Krogh in 1973.

The method just described proved to be a major technical breakthrough as it resulted in a reduction in lead-background contamination by a factor of between 10,000 and nearly 1,000,000. This means that a single grain can now be analyzed with a lower contamination level (or background correction) than was possible before with 100,000 similar grains. Advances in high-sensitivity mass spectrometry of course were essential to this development.

Ion-exchange chromatography

Once dissolved, the sample is ready for the chemical separation of the dating elements. This is generally achieved by using the methods of ion-exchange chromatography. In this process, ions are variously adsorbed from solution onto materials with ionic charges on their surface and separated from the rest of the sample. After the dating elements have been isolated, they are loaded into a mass spectrometer and their relative isotopic abundances determined.

The abundance of certain isotopes used for dating is determined by counting the number of disintegrations per minute (*i.e.*, emission activity). The rate is related to the number of such atoms present through the half-life (see above). For example, a certain amount of carbon-14 (^{14}C) is present in all biological components at the Earth's surface. This radioactive carbon is continually formed when nitrogen atoms of the upper atmosphere collide with neutrons produced by the interaction of high-energy cosmic rays with the atmosphere. An organism takes in small amounts of carbon-14, together with the stable (nonradioactive) isotopes carbon-12 (^{12}C) and carbon-13 (^{13}C), as long as it is alive. Once it dies, however, no additional carbon-14 is acquired and the level of radiocarbon in the organism's tissue decreases progressively as a function of half-life. The time that has passed since the organism was alive can be determined by counting the beta emissions from a tissue sample. The number of emissions in a given time period is proportional to the amount of residual carbon-14.

The introduction of an instrument called an accelerator mass spectrometer has brought about a major advance in radiocarbon dating. Unlike the old detector (*e.g.*, the Geiger counter) that counts the few decay particles emitted from a large amount of carbon, the new instrument counts directly all of the carbon-14 atoms in a sample. This increase in instrument sensitivity has made it possible to reduce the sample size by as much as 10,000 times and at the same time improve the precision of ages measured. (For a detailed discussion of radiocarbon age determination, see below *Carbon-14 dating and other cosmogenic methods.*)

In situ analysis of mineral grains

In a similar development, the use of highly sensitive thermal ionization mass spectrometers is replacing the counting techniques employed in some disequilibrium dating (see below). Not only has this led to a reduction in sample size and measurement errors but it also has permitted a whole new range of problems to be investigated. Certain parent-daughter isotopes are extremely refractory and do not ionize in a conventional mass spectrometer. To solve this problem, researchers are developing new instruments in which a small amount of material can be evaporated from the surface with a pulse of energy and ionized with a pulse of laser light. A major trend anticipated in geochronology and isotope geochemistry involves the analysis of mineral grains in place without chemical dissolution and mass spectrometry. This type of analysis requires expensive equipment in which a focused beam of ions is directed at a spot on a mineral sample. This causes atoms to evaporate from the surface, and the ions produced are extracted and measured in a mass spectrometer. Uranium-lead dating of zircon by this method has been pioneered by William Compston at the Australian National University.

MAJOR METHODS OF ISOTOPIC DATING

Isotopic dating relative to fossil dating requires a great deal of effort and depends on the integrated specialized skills of geologists, chemists, and physicists. It is, nevertheless, a valuable resource that allows correlations to be made over virtually all of Earth history with a precision once only possible with fossiliferous units that are restricted to the most recent 12 percent or so of geologic time. Although any method may be attempted on any unit, the best use of this resource requires that every effort be made to tackle each problem with the most efficient technique. Because of the long half-life of some isotopic systems or the high background or restricted range of parent abundances, some methods are inherently more precise. The skill of a geochronologist is demonstrated by the ability to attain the knowledge required and the precision necessary with the least number of analyses. The factors considered in selecting a particular approach are explored here.

Uranium–lead method. As each dating method was developed, tested, and improved, mainly since 1950, a vast body of knowledge about the behaviour of different isotopic systems under different geologic conditions has evolved. It is now clear that with recent advances the uranium–lead method is superior in providing precise age information with the least number of assumptions. The method has evolved mainly around the mineral zircon ($ZrSiO_4$). Because of the limited occurrence of this mineral, it was once true that only certain felsic igneous rocks (those consisting largely of the light-coloured, silicon and aluminum-rich minerals feldspar and quartz) could be dated. Today, however, baddeleyite (ZrO_2) has been found to be widespread in the silica-poor mafic igneous rocks (see above). In addition, perovskite ($CaTiO_3$), a common constituent of some ultramafic igneous rocks, has been shown to be amenable to precise uranium–lead dating. As a result of these developments, virtually all igneous rocks can now be dated. This capability, moreover, has been enhanced because the most advanced geochronological laboratories are able to analyze samples that weigh only a few millionths of a gram. This amount can be found in a comparatively large number of rocks, whereas the amount previously required (about 0.1 gram) cannot. Age determinations also can now be made of low-uranium trace minerals such as rutile (TiO_2), a common constituent found in mineral deposits, adding still further to the number of entities that are datable by the uranium–lead method. Other minerals commonly employed to date igneous and metamorphic rocks include titanite, monazite, and even garnet in certain favourable cases. Additional minerals listed in Table 2 have been tried with varying success.

Double uranium–lead chronometers

The reason why uranium–lead dating is superior to other methods is simple: there are two uranium–lead chronometers. Because there exist two radioactive uranium atoms (those of mass 235 and 238), two uranium–lead ages can be calculated for every analysis. The age results or equivalent daughter–parent ratio can then be plotted one against the other on a concordia diagram, as shown in Figure 5. If the point falls on the upper curve shown, the locus of identical ages, the result is said to be concordant, and a closed-system unequivocal age has been established. Any leakage of daughter isotopes from the system will cause the two ages calculated to differ, and data will plot below the curve. Because each of the daughters has a different half-life, early leakage will affect one system more than the other. Thus there is a built-in mechanism that can prove or disprove whether a valid age has been measured. Historically it had been observed that the uranium–lead systems in the mineral zircon from unmetamorphosed rocks were almost invariably disturbed or discordant but yielded a linear array on the concordia diagram. Given a set of variably disturbed samples, an extrapolation to zero

From T E Krogh, *Geochimica et Cosmochimica Acta*, vol. 46, © 1982 Pergamon Press

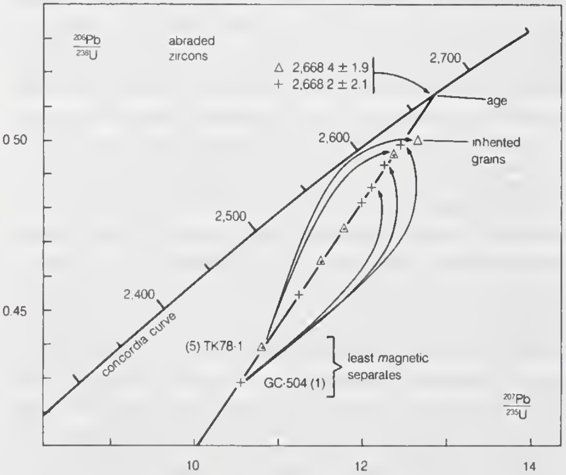


Figure 5: *Concordia diagram.*
Uranium–lead discordia line for isotopic results for zircons from two 2,668-million-year-old granites. Data for different fractions of zircon define a straight line with points plotting between 3 and 18 percent below the curve. The data for the least magnetic fractions represent the starting material in a series of tests designed to obtain data closer to the curve. The grains analyzed to obtain these data at the top of the arrows were selected to be crack-free, and all had natural surfaces removed by abrasion. These give the most reliable age information as extrapolations are reduced (see text).

disturbance was possible (see Figure 5). More recently, it has been found that of all the grains present in a rock a very few still retain closed isotopic systems but only in their interior parts. Thus grains with a diameter comparable to that of a human hair, selected under a microscope to be crack-free and of the highest possible quality, have been found to be more concordant than cracked grains. In addition, it has been shown that most such grains can be made much more concordant by mechanically removing their outer parts using an air-abrasion technique (upper points in Figure 5). Of course, the ability to analyze samples weighing only a few millionths of a gram was essential to this development. As noted earlier, this in turn was possible solely because the lead background contamination had been reduced from 1 × 10^{-6} grams to almost 1 × 10^{-12} grams per analysis. The methods of selection and abrasion used to locate grains with closed isotopic systems could be worked out only because the uranium–lead method has the inherent ability to assess with a single analysis whether or not a closed isotopic system has prevailed.

The presence of two radioactive parents provides a second major advantage because, as daughter products, lead atoms are formed at different rates and their relative abundance undergoes large changes as a function of time. Thus the ratio of lead-207 to lead-206 changes by about 0.1 percent every two million years. Since this ratio is easily calibrated and reproduced at such a level of precision, errors as low as ±2 million years at a confidence level of 95 percent are routinely obtained on lead-207–lead-206 ages. By contrast, errors as high as ±30 to 50 million years are usually quoted for the rubidium–strontium and samarium–neodymium isochron methods (see below).

Significance of zircon

The mineral zircon adds three more fundamental advantages to uranium–lead dating. First, its crystal structure allows a small amount of tetravalent uranium to substitute for zirconium, but excludes with great efficiency the incorporation of lead. (It might be said that one begins with an empty box.) Second, zircon, once formed, is highly resistant to change and has the highest blocking temperature ever observed. Finally, with few predictable exceptions, zircon grows or regrows only in liquid rock or in solid rock reheated to approach its melting point. Combining all of these attributes, it is often possible to measure both the time of crystallization and the time of second melting in different parts of the same grain or in different selected grains from the same rock. Of course, such a high blocking temperature can have its disadvantages. Inherited cores may give a mixed false age when the age of crystallization is sought. For this reason, three or more grain types or parts of a grain are analyzed to establish that material of only one age is present.

Experience with the results of the uranium–lead method for zircons has demonstrated an interesting paradox. If left at low surface temperatures for a geologically long time, the radioactivity within the crystal can destroy the crystal lattice structure, whereas at higher temperatures this process is self-annealing. In fact, when examined by X-ray methods, some zircons have no detectable structure, indicating that at least 25 percent of the initial atoms have been displaced by radiation damage. Under these conditions a low-temperature event insufficient to even reset the potassium–argon system (see below) in biotite can cause lead to be lost in some grains. It is no coincidence that, when criteria were finally found to locate concordant grains, these grains were also found to be those with the lowest uranium content and the lowest related radiation damage.

Determining the time of both the primary and secondary events

Given the two related uranium–lead parent–daughter systems, it is possible to determine both the time of the initial, or primary, rock-forming event and the time of a major reheating, or secondary, event. This is illustrated in Figure 6. Here, the uranium–lead isotopes in the mineral titanite ($CaTiSiO_5$) from a series of rocks that have a common geologic history plot on a straight line. The minerals first formed 1,651 million years ago but were later heated and lost varying amounts of lead 986 million years ago. In many cases, new titanite, distinguishable on the basis of colour, has formed in the same rock, while older, partly reset titanite is still present. Data points 14 and 7 in the figure represent such a pair. On this diagram, the presence of surface-correlated lead loss will displace data from the titanite line toward the time of loss close to zero age on the concordia curve. The uranium–lead data then would plot below the line shown, and neither the primary nor secondary age would be defined. The importance of eliminating recent loss, as discussed above, is clearly evident. It should be noted that, if these ages had been measured by any of the other schemes that have only a single parent–daughter pair, a whole series of different numbers spanning the time from 1.65 billion to 988 million years ago would be observed. There would be no way of telling which of the measured ages, if any, was valid.

Uranium–lead dating relies on the isolation of very high-quality grains or parts of mineral grains that are extremely rare but nevertheless present in most igneous, metamorphic, and sedimentary rock units. Samples weighing 10 to 50 kilograms are collected, crushed, and ground into a fine sand, and the various minerals are isolated on the basis of specific gravity, grain size, and magnetic properties. The minerals used are not visible in the field, but their presence can be inferred from the easily identified major minerals present.

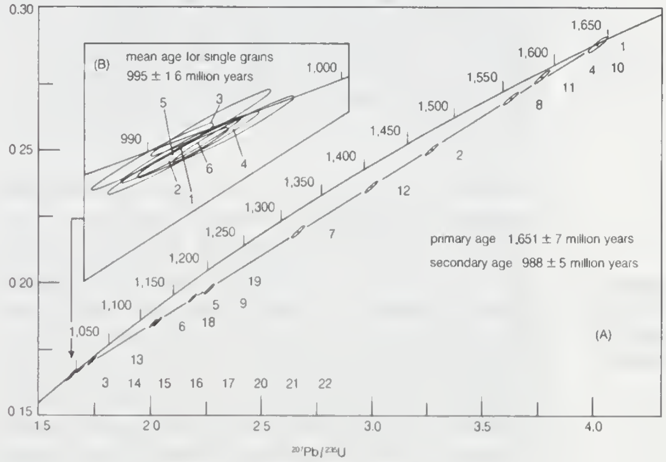


Figure 6: *Titanite discordia.*
(A) Uranium–lead isotopic data for a suite of titanite samples collected from a region in Labrador, Can., where rocks with titanite that initially formed 1,651 million years ago were subjected to a major heating event 988 million years ago. Variable amounts of lead lost by solid diffusion from the titanite from different samples at 988 million years produced the discordia line shown. Totally reset or newly grown grains plot at the lower end of the line. (B) The time of the second event also can be determined by dating zircon formed in rock melted during the second event. Replicate analysis of single zircon grains from this melt demonstrate the reproducibility of the method. The data average 0.4 percent discordant and indicate an uncertainty of ±3.5 million years for a single analysis and ±1.6 million years for the mean lead-207–lead-206 age of 995 million years at a confidence level of 95 percent. Grains analyzed weighed between 6 and 14 × 10^{-6} gram each.

One of the most interesting applications of the improved uranium–lead zircon technique has to do with its ability to achieve nearly concordant results from single grains extracted from sandstone. This is possible because zircon is chemically inert and is not disturbed during weathering and because single grains with a diameter about the thickness of a human hair contain sufficient uranium and lead for analysis in the most advanced laboratories. The examples given in Figure 7 indicate that a sandstone that underlies most of the province of Nova Scotia in Canada was probably originally deposited off the coast of North Africa and thrust over the continent before the opening of the Atlantic Ocean. This follows because the ages observed occur in North Africa, whereas those common in North America are absent.

Another sample, this one from sandstone deposited by a large river in northern Scotland (Figure 8), must have been derived from continental rocks whose ages are represented by those determined for the individually dated sand grains. In this case, the continent from which the sand was

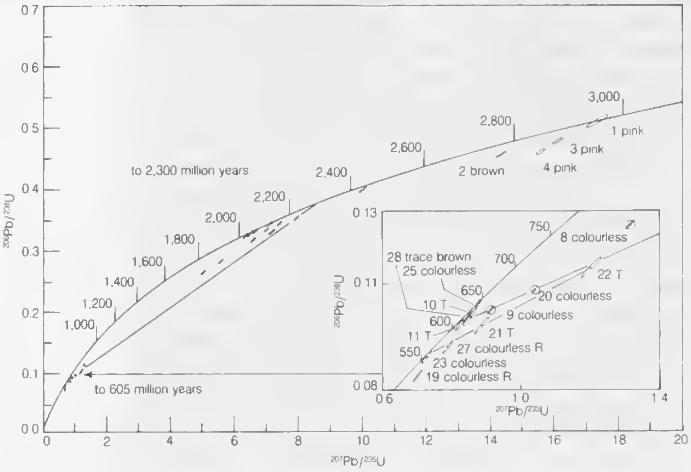


Figure 7: Uranium-lead results for single and multigrain zircon fractions from a type of sandstone that underlies most of Nova Scotia, Can. The cluster of ages around 600 million, 2,000 to 2,300 million, and 2,800 to 3,000 million years suggests that North Africa may have been the source of this sediment. It should be noted that the ages between 1,000 and 1,800 million years common in North America are absent (see text).

derived has moved away as a result of continental drift, but it can be identified by the ages measured.

Application of the isochron method

Rubidium–strontium method. The radioactive decay of rubidium-87 (^{87}Rb) to strontium-87 (^{87}Sr) was the first widely used dating system that utilized the isochron method. Rubidium is a relatively abundant trace element in the Earth's crust and can be found in many common rock-forming minerals in which it substitutes for the major element potassium. Because rubidium is concentrated in crustal rocks, the continents have a much higher abundance of the daughter isotope strontium-87 compared with the stable isotopes. This relative abundance is expressed as the ^{87}Sr/^{86}Sr ratio, where strontium-86 is chosen to represent the stable isotopes strontium-88, strontium-86, and strontium-84, which occur in constant proportions in natural materials. Thus a precise measurement of the ^{87}Sr/^{86}Sr ratio in a modern volcano can be used to determine age if recycled older crust is present. A ratio for average continental crust of about 0.72 has been determined by measuring strontium from clamshells from the major river systems. In contrast, the Earth's most abundant lava rocks, which represent the mantle and make up the major oceanic ridges, have values between 0.703 and 0.705. This difference may appear small, but considering that modern instruments can make the determination to a few parts in 70,000, it is quite significant. Dissolved strontium in the oceans today has a value of 0.709 that is dependent on the relative input from the continents and the ridges. In the geologic past changes in the activity of these two sources has produced varying ^{87}Sr/^{86}Sr ratios over time. Thus if well-dated, unaltered fossil shells containing strontium from ancient seawater are analyzed, changes in this ratio with time can be observed and applied in reverse to estimate the time when fossils of unknown age were deposited.

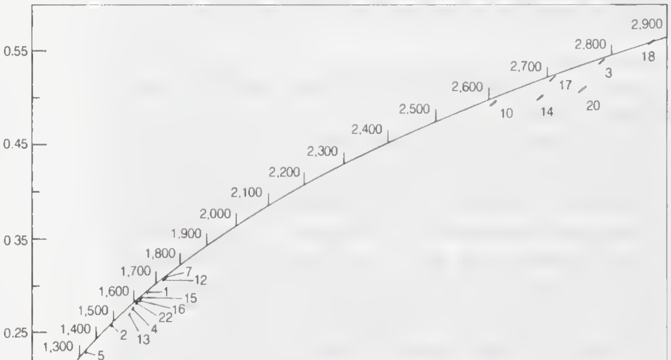


Figure 8: Uranium-lead age results for single zircon grains extracted from sandstone deposited by a river in northern Scotland. The source continent must have had rocks whose ages clustered around 1,100 to 1,320 million, 1,550 to 1,830 million, and 2,600 to 2,860 million years ago. Northern Labrador in Canada has rocks with such an age distribution and thus is a possible source for the sediment (see text).

Dating simple igneous rocks. The rubidium–strontium pair is ideally suited for the isochron dating of igneous rocks. As a liquid rock cools, first one mineral and then another achieves saturation and precipitates, each extracting specific elements in the process. Strontium is extracted in many minerals that are formed early, whereas rubidium is gradually concentrated in the final liquid phase. At the time of crystallization, this produces a wide range in the Rb/Sr ratio in rocks that have identical ^{87}Sr/^{86}Sr ratios. On the isochron diagram shown in Figure 4 above, the samples would plot initially at points R_1 to R_3 along a line representing the initial ratio designated $(^{87}Sr/^{86}Sr)_0$.

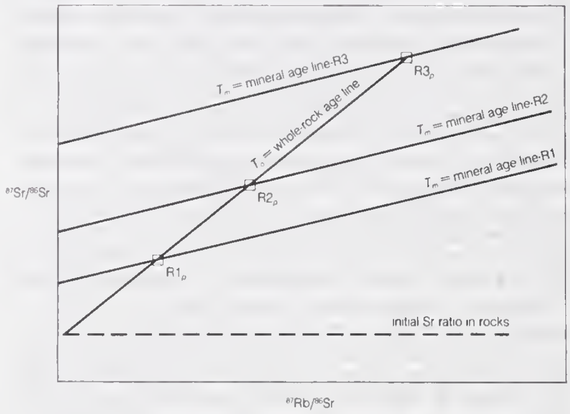

T_m = mineral age line-R3
$R3_p$
T_0 = whole-rock age line
T_m = mineral age line-R2
T_m = mineral age line-R1
$R2_p$
$R1_p$
$^{87}Sr/^{86}Sr$
initial Sr ratio in rocks
$^{87}Rb/^{86}Sr$

Figure 9: *Rubidium–strontium isochron diagram.* Hypothetical diagram for three rock samples with present-day isotopic values that plot at $R1_p$, $R2_p$, and $R3_p$ and define the whole rock age (T_0) and initial ratio shown. Minerals within each rock exchanged strontium isotopes during metamorphism, so that they have achieved identical initial ratios. Exchange between rocks has not taken place, however; thus, data for each rock plot on a primary isochron, while data for minerals define a series of parallel isochrons that indicate the time since metamorphism.

Over geologic time, this ratio is increased in proportion to the ^{87}Rb/^{86}Sr ratio, as discussed earlier, and the line rotates with a slope equal to $(e^{\lambda t} - 1)$ that represents the time elapsed; thus, the present-day ratio $(^{87}Sr/^{86}Sr)_p$ equals the initial ratio $(^{87}Sr/^{86}Sr)_0$ plus radiogenic additions, or $(^{87}Sr/^{86}Sr)_p = (^{87}Sr/^{86}Sr)_0 + {}^{87}Rb/^{86}Sr\ (e^{\lambda t} - 1)$. This equation is that of a straight line of the form $y = b + xm$, where $y = (^{87}Sr/^{86}Sr)_p$, the value measured today; b represents $(^{87}Sr/^{86}Sr)_0$, the value initially present; x stands for the ^{87}Rb/^{86}Sr ratio; and m is the slope of the line $(e^{\lambda t} - 1)$.

In practice, rock samples weighing several kilograms each are collected from a suite of rocks that are believed to have been part of a single homogeneous liquid prior to solidification. The samples are crushed and homogenized to produce a fine representative rock powder from which a fraction of a gram is withdrawn and dissolved in the presence of appropriate isotopic traces, or spikes. Strontium and rubidium are extracted and loaded into the mass spectrometer, and the values appropriate to the x and y coordinates are calculated from the isotopic ratios measured. Once plotted as $R1_p$ (*i.e.*, rock 1 present values), $R2_p$, and $R3_p$, as in Figure 9, the data are examined to assess how well they fit the required straight line. Using estimates of measurement precision, the crucial question of whether or not scatter outside of measurement error exists is addressed. Such scatter would constitute a geologic component, indicating that one or more of the underlying assumptions has been violated and that the age indicated is probably not valid. For an isochron to be valid, each sample tested must (1) have had the same initial ratio,

(2) have been a closed system over geologic time, and (3) have the same age.

Well-preserved, unweathered rocks that crystallized rapidly and have not been subjected to major reheating events are most likely to give valid isochrons. Weathering is a disturbing influence, as is leaching or exchange by hot crustal fluids, since many secondary minerals contain rubidium. Volcanic rocks are most susceptible to such changes because their minerals are fine-grained and unstable glass may be present. On the other hand, meteorites that have spent most of their time in the deep freeze of outer space can provide ideal samples.

Suitability of potassium-bearing minerals for rubidium–strontium dating

Dating minerals. Potassium-bearing minerals including several varieties of mica, are ideal for rubidium–strontium dating as they have abundant parent rubidium and a low abundance of initial strontium. In most cases, the changes in the $^{87}Sr/^{86}Sr$ ratio are so large that an initial value can be assumed without jeopardizing the accuracy of the results. Considered in terms of Figure 9, this would mean that the slope is determined from a single point that plots so far to the right and so far above the diagram that the actual initial value on the y axis can be easily estimated. When minerals with a low-rubidium or a high-strontium content are analyzed, the isochron-diagram approach can be used to provide an evaluation of the data. As discussed above, rubidium–strontium mineral ages need not be identical in a rock with a complex thermal history, so that results may be meaningful in terms of dating the last heating event but not in terms of the actual age of a rock.

Dating metamorphic rocks. Should a simple igneous body be subjected to an episode of heating or of deformation or of a combination of both, a well-documented special data pattern develops. With heat, daughter isotopes diffuse out of their host minerals but are incorporated into other minerals in the rock. Eventually the $^{87}Sr/^{86}Sr$ ratio in the minerals becomes identical. When the rock again cools, the minerals close and again accumulate daughter products to record the time since the second event. Remarkably, the isotopes remain within the rock sample analyzed, and so a suite of whole rocks can still provide a valid primary age. This situation is easily visualized on an isochron diagram such as that in Figure 9, where a series of rocks plots on a steep line showing the primary age, but the minerals in each rock plot on a series of parallel lines that indicate the time since the heating event. If cooling is very slow, the minerals with the lowest blocking temperature, such as biotite mica, will fall below the upper end of the line.

A more dramatic presentation of this phenomenon is found when the changes in the $^{87}Sr/^{86}Sr$ ratios in a variety of minerals in a single rock are depicted as a function of geologic time, as in Figure 10. Here, an essentially rubidium-free, strontium-rich phase like apatite retains its initial $^{87}Sr/^{86}Sr$ ratio over time, whereas the value in such rubidium-rich, strontium-poor minerals as biotite increases rapidly with time. The rock itself gives the integrated, more gradual increase. At the time of heating, identical $^{87}Sr/^{86}Sr$ ratios are again achieved as described above, only to be followed by a second episode of isotopic divergence.

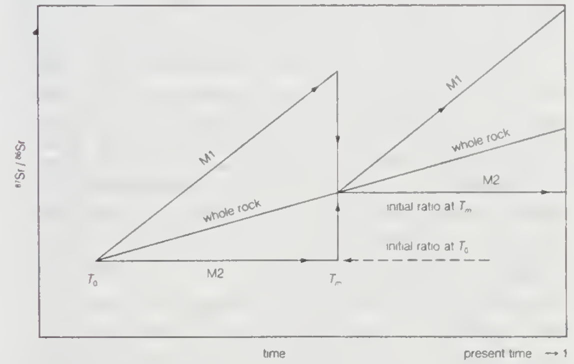


Figure 10: *Rubidium–strontium evolution diagram.* Isotopic changes in different minerals from the same rock are shown here as a function of time. Minerals formed at T_0 develop different isotopic ratios in proportion to their $^{87}Rb/^{86}Sr$ ratios, which are represented as the slope. Isotopic homogenization occurs at the time of metamorphism (T_m), but the isotopic ratios in different minerals again diverge toward their present values. Mineral M1 represents a rubidium-rich, strontium-poor mineral such as mica, while M2 represents a rubidium-poor, strontium-rich phase like apatite or plagioclase. The rock-growth line is the integrated result for all minerals present (see text).

Approaches to this ideal case are commonly observed, but peculiar results are found in situations where the heating is minimal. If one assumes for a moment that only the mineral with the lowest blocking temperature loses its daughter isotope, it is easy to imagine that other low-temperature minerals formed at this time may acquire extremely high $^{87}Sr/^{86}Sr$ ratios. Epidote, a low-temperature alteration mineral with a very high concentration of radiogenic strontium, has been found in rocks wherein biotite has lost strontium by diffusion. The rock itself has a much lower ratio, so that it did not take part in this exchange.

Impact of the rubidium–strontium method

Although rubidium–strontium dating is not as precise as the uranium–lead method, it was the first to be exploited and has provided much of the prevailing knowledge of Earth history. The procedures of sample preparation, chemical separation, and mass spectrometry are relatively easy to carry out, and datable minerals occur in most rocks. Precise ages can be obtained on high-level rocks (*i.e.,* those closer to the surface) and meteorites, and imprecise but nevertheless valuable ages can be determined for rocks that have been strongly heated. The mobility of rubidium in deep-level crustal fluids and melts that can infiltrate other rocks during metamorphism as well as in fluids involved in weathering can complicate the results.

Samarium–neodymium method. The radioactive decay of samarium of mass 147 (^{147}Sm) to neodymium of mass 143 (^{143}Nd) has been shown to be capable of providing useful isochron ages for certain geologic materials. Both parent and daughter belong to the rare earth element group, which is itself the subject of numerous geologic investigations. All members of this group have similar chemical properties and charge, but differ significantly in size. Because of this, they are selectively removed as different minerals are precipitated from a melt. In the opposite sense, their relative abundance in a melt can indicate the presence of certain residual minerals during partial melting. Unlike rubidium, which is enriched over strontium in the crust, samarium is relatively enriched with respect to neodymium in the mantle. Consequently, a volcanic rock composed of melted crust would have elevated radiogenic strontium values and depressed radiogenic neodymium values with respect to the mantle. As a parent–daughter pair, samarium-147 and neodymium-143 are unique in that both have very similar chemical properties, and so loss by diffusion may be reduced. Their low concentrations in surface waters indicates that changes during low-temperature alteration and weathering are less likely. Their presence in certain minerals in water-deposited gold veins, however, does suggest mobility under certain conditions. In addition, their behaviour under high-temperature metamorphic conditions is as yet poorly documented.

The exploitation of the samarium–neodymium pair for dating only became possible when several technical difficulties were overcome. Procedures to separate these very similar elements and methods of measuring neodymium isotope ratios with uncertainties of only a few parts in 100,000 had to be developed.

Dating igneous rocks

In theory, the samarium–neodymium method is identical to the rubidium–strontium approach (see above). Both use the isochron method to display and evaluate data. In the case of samarium–neodymium dating, however, the chemical similarity of parent and daughter adds another complication because fractionation during crystallization is extremely limited. This makes the isochrons short and adds further to the necessity for high precision. The result is that, with few exceptions, uncertainties in measured ages are as large as 30 to 100 million years, even for the oldest rocks and meteorites. Mineral isochrons provide the best results.

The equation relating present-day neodymium isotopic abundance as the sum of the initial ratios and radiogenic additions is that of a straight line, as discussed earlier for rubidium–strontium.

Figure 11 shows the successful application of the samarium–neodymium method to a sample of basalt from the Moon. Here, the constituent minerals plagioclase, ilmenite, and pyroxene provide enough spread in the ^{147}Sm/^{143}Nd ratio to allow an age of 3,700 ± 70 million years to be calculated. Other successful examples have been reported where rocks with open rubidium–strontium systems have been shown to have closed samarium–neodymium systems. In other examples, the ages of rocks with insufficient rubidium for dating have been successfully determined. There is considerable promise for dating garnet, a common metamorphic mineral, because it is known to concentrate the parent isotope.

From Geochimica et Cosmochimica Acta, Proceedings of the 6th Lunar Science Conference, 1975

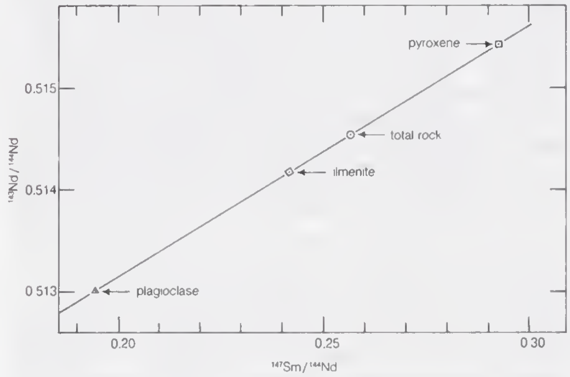


Figure 11: *Samarium–neodymium isochron for lunar basalt.* In this samarium–neodymium evolution diagram for a sample of medium-grained lunar basalt returned by the Apollo 17 astronauts, the data points for the total rock and for the plagioclase, ilmenite, and pyroxene mineral separates form a precise linear array. The "best-fit" line through these points represents a mineral isochron and yields a crystallization age of 3,700 ± 70 million years at a confidence level of 95 percent. The initial ^{143}Nd/^{144}Nd ratio is 0.50825 ± 12.

Dating meteorites and lunar rocks

In general, the use of the samarium–neodymium method as a dating tool is limited by the fact that other methods (mainly the uranium–lead approach) are more precise and require fewer analyses. In the case of meteorites and lunar rocks where samples are limited and minerals for other dating methods are not available, the samarium–neodymium method can provide the best ages possible.

Rhenium–osmium method. The decay scheme in which rhenium-187 is transformed to osmium-187 shows promise as a means of studying mantle–crust evolution but has displayed only limited potential for isotopic dating. Technical difficulties have yet to be overcome. Osmium is strongly concentrated in the mantle and extremely depleted in the crust, so that crustal osmium must have exceedingly high radiogenic-to-stable ratios while the mantle values are low. In fact, crustal levels are so low that they are extremely difficult to measure with current technology. Most work to date has centred around rhenium- or osmium-enriched minerals. Another problem that is still under investigation involves the difficulty of ionizing osmium in a mass spectrometer. A number of new approaches are being studied. In one method, success has been demonstrated in ionizing osmium in an instrument called an ion microprobe. This instrument bombards the sample target with a beam of ions of such high energy that atoms in the sample evaporate into a vacuum and a relatively small number becomes ionized. The ions are then extracted into a mass spectrometer, and the desired isotopic abundances are determined. Another related approach being tested involves the evaporation of the sample by a short pulse of ions, with a subsequent pulse of laser light tuned to ionize the desired element. Once the technical difficulties are surmounted, the distribution of rhenium and radiogenic and stable osmium isotopes will be explored. The greatest potential for this method might be in studies concerning the origin and age of mineral deposits since rhenium and osmium are known to occur in such materials.

Use of an ion microprobe

Potassium–argon methods. The radioactive decay scheme involving the breakdown of potassium of mass 40 (^{40}K) to argon gas of mass 40 (^{40}Ar) formed the basis of the first widely used isotopic dating method. Since radiogenic argon-40 was first detected in 1938 by the American geophysicist Lyman T. Aldrich and A.O. Nier, the method has evolved into one of the most versatile and widely employed methods available. Potassium is one of the 10 most abundant elements that together make up 99 percent of the Earth's crust and is therefore a major constituent of many rock-forming minerals. In fact, potassium-40 decays to both argon-40 and calcium-40, but because argon is absent in most minerals while calcium is present, the argon produced is easier to detect and measure. Most of the argon in the Earth's atmosphere has been created by the decay of potassium-40 as the argon-40 abundance is about 1,000 times higher than expected from cosmic abundances. Argon dating involves a different technology from all the other methods so far described because argon exists as a gas at room temperature. Thus it can be purified as it passes down a vacuum line by freezing out or reacting out certain contaminants. It is then introduced into a mass spectrometer through a series of manual or computer-controlled valves. Technical advances, including the introduction of the argon-40–argon-39 method and laser heating, that have improved the versatility of the method, are described below.

In conventional potassium–argon dating, a potassium-bearing sample is split into two fractions: one is analyzed for its potassium content, while the other is fused in a vacuum to release the argon gas. After purification has been completed, a spike enriched in argon-38 is mixed in and the atomic abundance of the daughter product argon-40 is measured relative to the argon-38 added. The amount of the argon-36 present is then determined relative to argon-38 to provide an estimate of the background atmospheric correction. In this case, relatively large samples, which may include significant amounts of alteration, are analyzed. Since potassium is usually added by alteration, the daughter–parent ratio and the age might be too low.

Argon-40–argon-39 method

A method designed to avoid such complexities was introduced by the geochronologists Craig M. Merrihue and Grenville Turner in 1966. In this technique, known as the argon-40–argon-39 method, both parent and daughter can be determined in the mass spectrometer as some of the potassium atoms in the sample are first converted to argon-39 in a nuclear reactor. In this way, the problem of measuring the potassium in inhomogeneous samples is eliminated and smaller amounts of material can be analyzed. An additional advantage then becomes possible. The sample can be heated in stages at different temperatures and the age calculated at each step. If alteration is evident, the invalid low-temperature age can be eliminated and a valid high-temperature age determined. In some cases, partly reset systems also may be detected.

As in all dating systems, the ages calculated can be affected by the presence of inherited daughter products. In a few cases, argon ages older than that of the Earth which violate local relative age patterns have even been determined for the mineral biotite. Such situations occur mainly where old rocks have been locally heated, which released argon-40 into pore spaces at the same time that new minerals grew. Under favourable circumstances the isochron method may be helpful, but tests by other techniques may be required. For example, the rubidium–strontium method would give a valid isotopic age of the biotite sample with inherited argon.

As techniques evolved, argon background levels have been reduced and the method has become more and more sensitive. Capitalizing on this, it is now possible to measure the minute amount of argon released when a single spot on a crystal is heated by an intense laser beam. For geologically old potassium-rich materials, a single spot may produce sufficient gas for analysis, whereas single millimetre-sized grains may be required in very young materials. Progressive refinement of the method has made new areas of research possible, and the ability to understand complexities encountered in earlier investigations has increased. In one study the age of volcanic ash as young as 215,000 ± 4,000 years and the presence of inherited older grains in another ash sample were thoroughly documented. This was done by melting single millimetre-sized grains with a laser and measuring individual argon-40–argon-39 ages with a highly sensitive gas mass spectrometer. In Figure 12A an age line between present atmospheric argon ratios and that of the volcanic ash source dated at 215,000 ± 4,000 years old is defined. Figure 12B shows that another sample of similar volcanic ash yielded comparable data, but many grains with older age components that plot below the line are present.

From P. van den Bogaard *et al*, reprinted by permission from *Nature*, vol. 342, pp 523-24, copyright © 1989 Macmillan Magazines Ltd

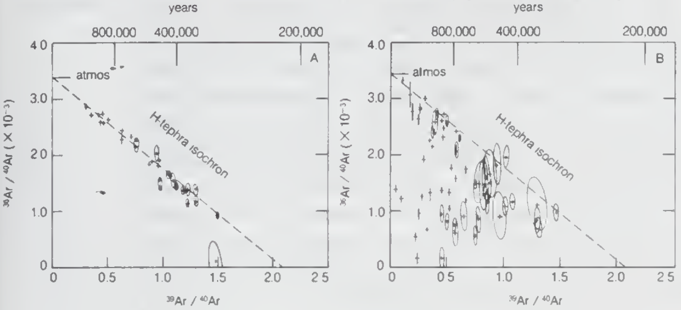


Figure 12: Potassium–argon isotope correlation diagram based on data for one of the youngest rocks ever dated by the potassium–argon method. In (A), data for single millimetre-sized feldspar grains plot on a line between the values for present-day atmospheric argon (atmos) and argon from a source 215,000 ± 4,000 years old (error quoted is at the 66-percent confidence level). Data for the same mineral from a different volcanic ash, shown in (B), indicate a similar age, but in this case many data points plot below the age line, suggesting that much older grains are present as well.

The potassium–argon method has provided a great deal of information about the Earth's recent and ancient past. It has been instrumental, for example, in determining the ages of the stripes of alternating normally and reversely magnetized volcanic rocks that parallel the axis of the mid-oceanic ridges. In ancient shield areas large segments of crust that were uplifted and cooled at the same time—*i.e.*, geologic provinces—have been identified by the potassium–argon method. The technique is highly responsive to thermal events in a relatively predictable fashion, so the cooling history of a region may be established. (T.E.Kr.)

Fission-track dating. This is a special type of dating method that makes use of a microscope rather than a mass spectrometer, and capitalizes on damaged zones, or tracks, created in crystals during the spontaneous fission of uranium-238. In this unique type of radioactive decay, the nucleus of a single parent uranium atom splits into two fragments of similar mass with such force that a trail of crystal damage is left in the mineral. Immersing the sample in an etching solution of strong acid or base enlarges the fission tracks into tube-shaped holes large enough to be seen under a high-powered microscope. The number of tracks present can be used to calculate the age of the sample if the uranium content is known. Fortunately the uranium content of precisely the spot under scrutiny can be obtained by a similar process when working with a polished crystal surface. The sample is bombarded with slow (thermal) neutrons in a nuclear reactor, resulting in induced fission of uranium-235 (as opposed to spontaneous fission of uranium-238). The fission tracks produced by this process are recorded by a thin plastic film placed against the surface of the sample. The uranium content of the material can then be calculated so long as the neutron dose is known. The age of the sample is obtained using the equation, age $= N \times Ds/Di \times 6 \times 10^{-8}$, in which N is the total neutron dose expressed as neutrons per square centimetre and Ds is the observed track density for spontaneous fission while Di is that for induced fission.

Spontaneous fission of uranium-238

The preservation of crystal damage (*i.e.*, the retention of fission tracks) is highly sensitive to temperature and varies from mineral to mineral. The technique can be used to determine mild thermal events as low as 100° C. Alternately, primary ages can be calculated if the rock was formed at the surface and cooled quickly. Under these conditions the calculated fission-track ages of two minerals with widely different annealing temperatures would be identical. The accuracy achieved depends on the number of tracks counted, so that artificial glass coloured with 10 percent uranium can be dated as soon as 30 years after manufacture. With uranium levels of a few parts per million, samples as young as 300,000 years can be dated by counting tracks for one hour. When dealing with very old materials, high-uranium samples must be avoided because there are so many interlocking tracks that they can no longer be counted.

A special feature of fission-track dating lies in its ability to map the uranium distribution within mineral grains. Figure 13 shows such a uranium map for single zircon grains. In this case, the outer zones that grew during a major melting event contain much more uranium than the grains originally present. The uranium–lead age would then be highly biased toward the younger event and the primary age could be determined only after the outer zones were removed. In practice, fission-track dates are regarded as cooling ages unless proved otherwise. It might also be noted that uncertainties in results may arise from an uneven distribution of uranium, statistical errors in counting, and inaccurate estimates of neutron flux (dose of neutrons).

Fission-track dating can be used on a wide variety of minerals found in most geologic materials, and it is relatively inexpensive to apply. Because closure temperatures vary widely from, say, 300° C for titanite and zircon to less than 100° C for biotite and apatite, valuable information can be obtained regarding the uplift and cooling rates of crustal rocks.

By courtesy of T.E. Krogh

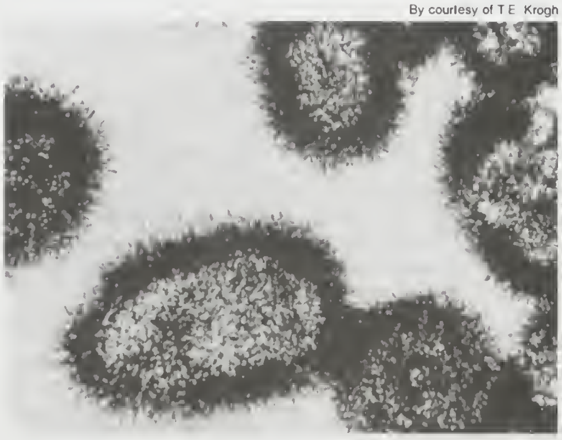


Figure 13: *Fission-track map.*
The uranium distribution within single zircon grains is recorded by a plastic film placed against a polished surface. The tiny black lines represent the etched path of crystal damage caused by the induced fission of uranium-235 in a nuclear reactor. The uranium in the zircon grains is concentrated in the outer rims.

Carbon-14 dating and other cosmogenic methods. The occurrence of natural radioactive carbon in the atmosphere provides a unique opportunity to date organic materials as old as 50,000 years. Unlike most isotopic dating methods, the conventional carbon-14 dating technique is not based on counting daughter isotopes. It relies instead on the progressive decay or disappearance of the radioactive parent with time.

The discovery of natural carbon-14 by Willard Libby of the United States began with his recognition that a process that had produced radiocarbon in the laboratory was also going on in the Earth's upper atmosphere—namely, the bombardment of nitrogen by free neutrons. Newly created carbon-14 atoms were presumed to react with atmospheric oxygen to form carbon dioxide (CO_2) molecules. Radioactive carbon thus was visualized as gaining entrance wherever atmospheric carbon dioxide enters—into land plants by photosynthesis, into animals that feed on the plants, into marine waters and freshwaters as a dissolved component, and from there into aquatic plants and animals. In short, all parts of the carbon cycle were seen to be invaded by the isotope carbon-14.

Discovery of carbon-14

Invasion is probably not the proper word for a component that Libby calculated should be present only to the extent of about one atom in a trillion stable carbon atoms. So low is such a carbon-14 level that no one had detected natural carbon-14 until Libby, guided by his own predictions, set out specifically to measure it. His success initiated a series of measurements designed to answer two questions: Is the concentration of carbon-14 uniform throughout the plant and animal kingdoms? And, if so, has today's uniform level prevailed throughout the recent past?

After showing the essential uniformity of carbon-14 in living material, Libby sought to answer the second question by measuring the radiocarbon level in organic samples dated historically—materials as old as 5,000 years from sources such as Egyptian tombs. With correction for radioactive decay during the intervening years, such old samples hopefully would show the same starting carbon-14 level as exists today. This was just what Libby's measurements indicated. His conclusion was that over the past 5,000 years the carbon-14 level in living materials has remained constant within the 5 percent precision of measurement. A dating method was thus available, subject only to confirmation by actual application to specific chronologic problems.

Since Libby's foundational studies, tens of thousands of carbon-14 measurements of natural materials have been made. Expressed as a fraction of the contemporary level, they have been mathematically converted to ages through equation 5 above. Archaeology has been the chief beneficiary of radioactive-carbon dating, but late glacial and postglacial chronological studies in geology have also been aided greatly.

Variations of the carbon-14 level in nature

Improvements in measurement accuracy and the ever-mounting experience in applying carbon-14 dating have provided superior and more voluminous data with which to better answer Libby's original questions. It is now clear that carbon-14 is not homogeneously distributed among today's plants and animals. The occasional exceptions all involve nonatmospheric contributions of carbon-14-depleted carbon dioxide to organic synthesis. Specifically, volcanic carbon dioxide is known to depress the carbon-14 level of nearby vegetation and dissolved limestone carbonate occasionally has a similar effect on freshwater mollusks, as does upwelling of deep ocean water on marine mollusks. In every case, the living material affected gives the appearance of built-in age.

In addition to spatial variations of the carbon-14 level, the question of temporal variation has received much study. A 2 to 3 percent depression of the atmospheric radioactive-carbon level since 1900 was noted soon after Libby's pioneering work, almost certainly the result of the dumping of huge volumes of carbon-14-free carbon dioxide into the air through smokestacks. Of more recent date was the overcompensating effect of man-made carbon-14 injected into the atmosphere during nuclear-bomb testing. The result was a rise in the atmospheric carbon-14 level by more then 50 percent. Fortunately, neither effect has been significant in the case of older samples submitted for carbon-14 dating. The ultimate cause of carbon-14 variations with time is generally attributed to temporal fluctuations in the cosmic rays that bombard the upper atmosphere and create terrestrial carbon-14. Whenever the number of cosmic rays in the atmosphere is low, the rate of carbon-14 production is correspondingly low, resulting in a decrease of the radioisotope in the carbon-exchange reservoir described above. Studies have revealed that the atmospheric radiocarbon level prior to 1000 BC deviates measurably from the contemporary level. In the year 6200 BC it was about 8 percent above what it is today. In the context of carbon-14 dating, this departure from the present-day level means that samples with a true age of 8,200 years would be dated by radiocarbon as 7,500 years old.

The problems stemming from temporal variations can be overcome to a large degree by the use of calibration curves in which the carbon-14 content of the sample being dated is plotted against that of objects of known age. In this way, the deviations can be compensated for and the carbon-14 age of the sample converted to a much more precise date. Calibration curves have been constructed using dendrochronological data (tree-ring measurements of bristlecone pines as old as 8,200 years); periglacial varve, or lake sediment, data (see above); and, in archaeological research, certain materials of historically established ages. It is clear that carbon-14 dates lack the accuracy that traditional historians would like to have. There may come a time when all radiocarbon ages rest on firmer knowledge of the sample's original carbon-14 level than is now available. Until then, the inherent error from this uncertainty must be recognized.

Carbon-14 contamination

A final problem of importance in carbon-14 dating is the matter of sample contamination. If a sample of buried wood is impregnated with modern rootlets or a piece of porous bone has recent calcium carbonate precipitated in its pores, failure to remove the contamination will result in a carbon-14 age between that of the sample and that of its contaminant. Consequently, numerous techniques for contaminant removal have been developed. Among them are the removal of humic acids from charcoal and the isolation of cellulose from wood and collagen from bone. Today, contamination as a source of error in samples younger than 25,000 years is relatively rare. Beyond that age, however, the fraction of contaminant needed to have measurable effect is quite small, and, therefore, undetected or unremoved contamination may occasionally be of significance.

A major breakthrough in carbon-14 dating occurred with the introduction of the accelerator mass spectrometer. This instrument is highly sensitive and allows precise ages on as little as one milligram of carbon, where the older method might require as much as 25 grams for ancient material. The increased sensitivity results from the fact that all of the carbon atoms of mass 14 can be counted in a mass spectrometer. By contrast, if carbon-14 is to be measured by its radioactivity, only those few atoms decaying during the measurement period are recorded. By using the accelerator mass spectrometer possible interference from nitrogen-14 is avoided since it does not form negative ion beams, and interfering molecules are destroyed by stripping electrons away by operating at several million volts.

Other cosmogenic methods

The development of the accelerator mass spectrometer has provided new opportunities to explore other rare isotopes produced by the bombardment of the Earth and meteorites by high-energy cosmic rays. Many of these isotopes have short half-lives and hence can be used to date events that happened in the past few thousand to a few million years. In one case, the time of exposure, like the removal of rock by a landslide, can be dated by the presence of the rare beryllium-10 (^{10}Be) isotope formed in the newly exposed surface of a terrestrial object or meteoroidal fragment by cosmic-ray bombardment. Other applications include dating groundwater with chlorine-36 (^{36}Cl), dating marine sediments with beryllium-11 (^{11}Be) and aluminum-26 (^{26}Al), and dating glacial ice with krypton-81 (^{81}Kr). In general, the application of such techniques is limited by the enormous cost of the equipment required.

Uranium-series disequilibrium dating. The isotopic dating methods discussed so far are all based on long-lived radioactive isotopes that have survived since the elements were created or on short-lived isotopes that were recently produced by cosmic-ray bombardment. The long-lived isotopes are difficult to use on young rocks because the extremely small amounts of daughter isotopes present are difficult to measure. A third source of radioactive isotopes is provided by the uranium- and thorium-decay chains. As noted in Table 3, these uranium–thorium series radioisotopes, like the cosmogenic isotopes, have short half-lives and are thus suitable for dating geologically young materials. The decay of uranium to lead is not achieved by a single step but rather involves a whole series of different elements, each with its own unique set of chemical properties.

In closed-system natural materials, all of these intermediate daughter elements exist in equilibrium amounts. That is to say, the amount of each such element present is constant and the number that form per unit time is identical to the number that decay per unit time. Accordingly, those with long half-lives are more abundant than those with short half-lives. Once a uranium-bearing mineral breaks down and dissolves, the elements present may behave differently and equilibrium is disrupted. For example, an isotope of thorium is normally in equilibrium with uranium-234 but is found to be virtually absent in modern corals even though uranium-234 is present. Over a long period of time uranium-234, however, decays to thorium-230, which results in a build-up of the latter in old corals and thereby provides a precise measure of time.

Most of the studies using the intermediate daughter elements were for years carried out by means of radioactive counting techniques—*i.e.,* the number of atoms present was estimated by the radioactivity of the sample. The introduction of highly sensitive mass spectrometers that allow the total number of atoms to be measured rather than the much smaller number that decay has resulted in a revolutionary change in the family of methods based on uranium and thorium disequilibrium.

Thorium-230 dating. The insoluble nature of thorium provides for an additional disequilibrium situation that allows sedimentation rates in the modern oceans to be determined. In this case, thorium-230 in seawater, produced principally by the decay of uranium-234, is deposited preferentially in the sediment without the uranium-234 parent. This is defined as excess thorium-230 because its abundance exceeds the equilibrium amount that should be present. With time, the excess decays away and the age of any horizon in a core sample can be estimated from the observed thorium-230-to-thorium-232 ratio in the seawater-derived component of the core. Sedimentation rates between 1 and 20 millimetres per 1,000 years are commonly found with slight variations between the major ocean basins.

Lead-210 dating. The presence of radon gas as a member of the uranium-decay scheme provides a unique method for creating disequilibrium. The gas radon-222 (^{222}Rn) escapes from the ground and decays rapidly in the atmosphere to lead-210 (^{210}Pb), which falls quickly to the surface where it is incorporated in glacial ice and sedimentary materials. By assuming that the present deposition rate also prevailed in the past, the age of a given sample at depth can be estimated by the residual amount of lead-210. (E.A.O./T.E.Kr.)

GEOLOGIC HISTORY OF THE EARTH

The geologic history of the Earth reveals much about the evolution of the continents, oceans, atmosphere, and biosphere. The layers of rock at the Earth's surface contain evidence of the evolutionary processes undergone by these components of the terrestrial environment during the times at which each layer was formed. By studying this rock record from the very beginning, it is thus possible to trace their development and the resultant changes through time.

THE PREGEOLOGIC PERIOD

The history of the Earth spans approximately 4.6 billion years. The oldest known rocks, however, have an isotopic age of only about 3.9 billion years. There is, in effect, a stretch of 700 million years for which no geologic record exists, and the evolution of this pregeologic period of time is not surprisingly the subject of much speculation. To understand this little-known period, the following factors have to be considered: the age of formation at 4.6 billion years ago, the processes in operation until 3.9 billion years ago, the bombardment of the Earth by meteorites, and the earliest zircon crystals.

It is widely accepted by both geologists and astronomers that the Earth is roughly 4.6 billion years old. This age has been obtained from the isotopic analysis of many meteorites as well as of soil and rock samples from the Moon by such dating methods as rubidium–strontium and uranium–lead (see above). It is taken to be the time when these bodies formed and, by inference, the time at which a significant part of the solar system developed. When the evolution of the isotopes of lead-207 and lead-206 is studied from several lead deposits of different age on Earth, including oceanic sediments that represent a homogenized sample of the Earth's lead, the growth curve of terrestrial lead can be calculated, and when this is extrapolated back in time it is found to coincide with the age of about 4.6 billion years measured on lead isotopes in meteorites. The Earth and meteorites thus have had similar lead-isotope histories, and so it is concluded that they condensed or accreted as solid bodies from a primeval cloud of interstellar gas and dust—the so-called solar nebula from which the entire solar system is thought to have formed—at about the same time.

Formation of the Earth

Particles in the solar nebula condensed to form solid grains, and with increasing electrostatic and gravitational influences they eventually clumped together into fragments or chunks of rock. One of these planetesimals developed into the Earth. The constituent metallic elements sank toward the centre of the mass, while lighter elements rose toward the top. The lightest ones (such as hydrogen and helium) that might have formed the first, or primordial, atmosphere probably escaped into outer space. In these earliest stages of terrestrial accretion heat was generated by three possible phenomena: (1) the decay of short-lived radioactive isotopes, (2) the gravitational energy released from the sinking of metals, or (3) the impact of small planetary bodies (or planetesimals). The increase in temperature became sufficient to heat the entire planet. Melting at depth produced liquids that were gravitationally light and thus rose toward the surface and crystallized to form the earliest crust. Meanwhile, heavier liquids rich in iron, nickel, and perhaps sulfur separated out and sank under gravity, giving rise to the core at the centre of the growing planet; and the lightest volatile elements were able to rise and escape by outgassing, which may have been associated with surface volcanic activity, to form the secondary atmosphere and the oceans. This chemical process of melting, separation of material, and outgassing is referred to as the differentiation of the Earth. The earliest thin crust was probably unstable and so foundered and collapsed to depth. This in turn generated more gravitational energy, which enabled a thicker, more stable, longer-lasting crust to form. Once the Earth's interior (or its mantle) was hot and liquid, it would have been subjected to large-scale convection, which may have enabled oceanic crust to develop above upwelling regions. Rapid recycling of crust–mantle material occurred in convection cells, and in this way the earliest terrestrial continents may have evolved during the 700-million-year gap between the formation of the Earth and the beginning of the rock record. It is known from direct observation that the surface of the Moon is covered with a multitude of meteorite craters. There are about 40 large basins attributable to meteorite impact. Known as maria, these depressions were filled in with basaltic lavas caused by the impact-induced melting of the lunar mantle. Many of these basalts have been analyzed isotopically and found to have crystallization ages of 3.9 to 4 billion years. It can be safely concluded that the Earth, with a greater attractive mass than the Moon, must have undergone more extensive meteorite bombardment. According to the English-born geologist Joseph V. Smith, a minimum of 500 to 1,000 impact basins were formed on the Earth within a period of about 100 to 200 million years prior to 3.95 billion years ago. Moreover, plausible calculations suggest that this estimate represents merely the tail end of an interval of declining meteorite bombardment and that about 20 times as many basins were formed in the preceding 300 million years. Such intense bombardment would have covered most of the Earth's surface, with the impacts causing considerable destruction of the terrestrial crust up to 3.9 billion years ago. There is, however, no direct evidence of this important phase of Earth history because rocks older than 3.9 billion years have not been preserved.

Differentiation of the Earth

An exciting discovery was made in 1983 by William Compston and his research group at the Australian National University with the aid of an ion microprobe (see above *Absolute dating: Instruments and procedures*). Compston and his associates found that a water-laid clastic sedimentary quartzite from Mount Narryer in western Australia contained detrital zircon grains that were 4.18 billion years old. In 1986 they further discovered that one zircon in a conglomerate only 60 kilometres away was 4.276 billion years old; 16 other grains were determined to be the same age or slightly younger. This is the oldest dated material on Earth. The rocks from which the zircons in the quartzites and conglomerates were derived have either disappeared or have not yet been found. The ages of these single zircon grains are significantly older than those of the oldest known intact rocks, which are granites discovered near the Great Slave Lake in northwestern Canada. The latter contain zircons that are 3.96 billion years old.

DEVELOPMENT OF THE ATMOSPHERE AND OCEANS

Formation of the secondary atmosphere. The Earth's secondary atmosphere began to develop at the time of planetary differentiation, probably in connection with volcanic activity. Its component gases, however, were most likely very different from those emitted by modern volcanoes. Accordingly, the composition of the early secondary atmosphere was quite distinct from that of today's atmosphere. Carbon monoxide, carbon dioxide, water vapour, and methane predominated; however, free oxygen could not have been present, since even modern volcanic gases contain no oxygen. It is therefore assumed that the secondary atmosphere during the Archean—the time of the oldest known rocks—was anoxygenic. The free oxygen that makes up the bulk of the present atmosphere evolved over geologic time by two possible processes. First, solar ultraviolet radiation (the short-wavelength component of sunlight) would have provided the energy needed to break up water vapour into hydrogen, which escaped into space, and free oxygen, which remained in the atmosphere. This process was in all likelihood important before the appearance of the oldest extant rocks, but after that time the second process, organic photosynthesis, became predominant. Primitive organisms, such as blue-green algae (or cyanobacteria), cause carbon dioxide and water to react by photosynthesis to produce carbohydrates, which they need for growth, repair, and other vital functions, and this reaction releases free oxygen. The discovery of stromatolites (layered or conical sedimentary structures formed by sediment-binding marine algae) in 3.5-billion-year-old limestones in several parts of the world indicates that blue-green algae existed by that time. The presence of such early carbonate sediments is evidence that carbon dioxide was present in the atmosphere, and it has been calculated that it was at least 100 times greater than the amount in the present-day atmosphere. It can be assumed that such abundant carbon dioxide would have caused retention of heat, resulting in a greenhouse effect and a hot atmosphere (see ATMOSPHERE).

What happened to all the oxygen that was released? It might be surprising to learn that it took at least 1 billion years before there was sufficient oxygen in the atmosphere for oxidative diagenesis to give rise to red beds (sandstones that are predominantly red in colour due to fully oxidized iron coating individual grains) and that 2.2 billion years passed before a large number of life-forms could evolve. An idea formulated by the American paleontologist Preston Cloud has been widely accepted as an answer to this question. The earliest primitive organisms produced free oxygen as a by-product, and in the absence of oxygen-mediating enzymes it was harmful to their living cells and had to be removed. Fortunately for the development of life on the early Earth there was extensive volcanic activity, which resulted in the deposition of much lava, the erosion of which released enormous quantities of iron into the oceans. This ferrous iron is water-soluble and therefore could be easily transported, but it had to be converted to ferric iron, which is highly insoluble, before it could be precipitated as iron formations. In short, the organisms produced the oxygen and the iron formations accepted it. Iron formations can be found in the earliest sediments (those deposited 3.8 billion years ago) at Isua in West Greenland, and thus this process must have been operative by this time. Early Precambrian iron formations are so thick and common that they provide the major source of the world's iron. Large quantities of iron continued to be deposited until about 2 billion years ago, after which time the formations decreased and disappeared from the sedimentary record. Sulfides also accepted oxygen in the early oceans to be deposited as sulfates in evaporites, but such rocks are easily destroyed. One finds, nonetheless, 3.5-billion-year-old barite/gypsum-bearing evaporites up to 15 metres thick and at least 25 kilometres in extent in the Pilbara region of Western Australia. It seems likely that the excess iron in the early oceans was finally cleared out by about 1.7 billion years ago, and this decrease in the deposition of iron formations resulted in an appreciable rise in the oxygen content of the atmosphere, which in turn enabled more eolian red beds to form. Further evidence of the lack of oxygen in the early atmosphere is provided by detrital uraninite and pyrite and by paleosols—*i.e.,* fossil soils. Detrital uraninite and pyrite are readily oxidized in the presence of oxygen and thus do not survive weathering processes during erosion, transport, and deposition in an oxygenous atmosphere. Yet, these minerals are well preserved in their original unoxidized state in conglomerates that have been dated to be more than 2.2 billion years old on several continents. Paleosols also provide valuable clues, as they were in equilibrium with the prevailing atmosphere. From analyses of early Precambrian paleosols it has been determined that the oxygen content of the atmosphere 2.2 billion years ago was one hundredth of the present atmospheric level (PAL).

Fossils of eukaryotes, which are organisms that require an oxygen content of about 0.02 PAL, bear witness to the beginning of oxidative metabolism. The first microscopic eukaryotes appeared about 1.4 billion years ago. Life-forms with soft parts, such as jellyfish and worms, developed in profusion, albeit locally, toward the end of the Precambrian about 650 million years ago, and it is estimated that this corresponds to an oxygen level of 0.1 PAL. By the time land plants first appeared, roughly 400 million years ago, atmospheric oxygen levels had reached their present values.

Development of the oceans. Volcanic degassing of volatiles, including water vapour, occurred during the early stages of crustal formation and gave rise to the atmosphere. When the surface of the Earth had cooled to below 100° C (212° F), the hot water vapour in the atmosphere would have condensed to form the early oceans. The existence of 3.5-billion-year-old stromatolites is, as noted above, evidence of the activity of blue-green algae, and this fact indicates that the Earth's surface must have cooled to below 100° C by this time. Also, the presence of pillow structures in basalts of this age attests to the fact that these lavas were extruded under water, and this probably occurred around volcanic islands in the early ocean. The abundance of volcanic rocks of Archean age (3.8 to 2.5 billion years ago) is indicative of the continuing role of intense volcanic degassing, but since the early Proterozoic (from 2.5 billion years ago), much less volcanic activity has occurred. Until about 2 billion years ago there was substantial deposition of iron formations, cherts, and various other chemical sediments, but from roughly that time onward the relative proportions of different types of sedimentary rock and their mineralogy and trace element compositions have been very similar to their Phanerozoic equivalents; it can be inferred from this relationship that the oceans achieved their modern chemical characteristics and sedimentation patterns from approximately 2 billion years ago (Figure 14). By the late Precambrian, some 1 billion years ago, ferric oxides were chemically precipitated, indicating the availability of free oxygen. During Phanerozoic time (the last 570 million years), the oceans have been steady-state chemical systems, continuously reacting with the minerals added to them via drainage from the continents and with volcanic gases at the oceanic ridges.

TIME SCALES

The geologic history of the Earth over nearly four billion years of time is surveyed in the remainder of this article.

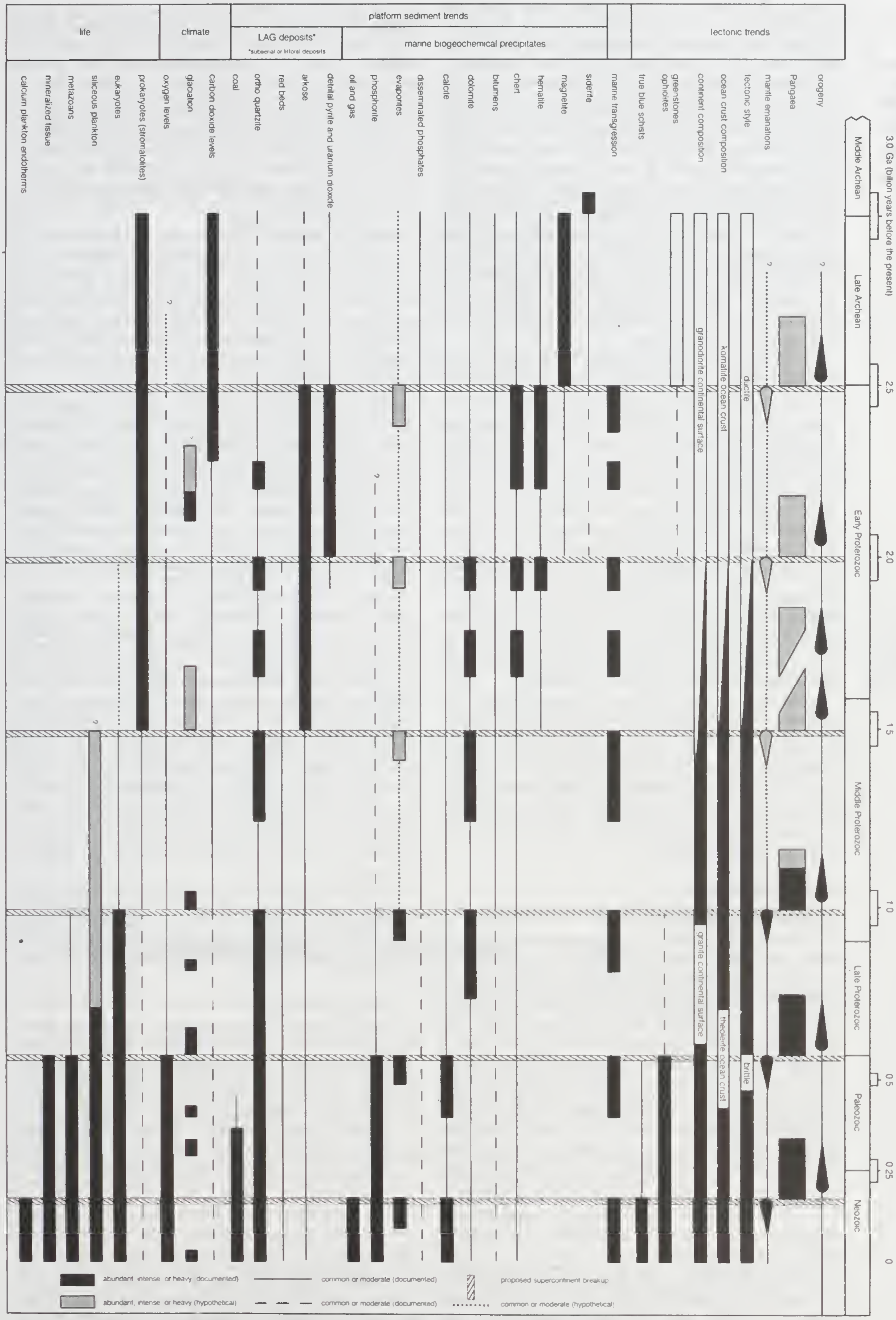


Figure 14: Qualitative summary of Late Archean to present tectonic, platform sediment, climatic, and biotic trends.

From T.R. Worsley, R.D. Nance, and J.B. Moody, *Geology*, vol. 14 (June 1986)

Chronostratigraphic and chronometric scales

Different types of phenomena and events in widely separated parts of the world have been correlated using an internationally acceptable, standardized time scale. There are, in fact, two geologic time scales. One is relative, or chronostratigraphic, and the other is absolute, or chronometric. The chronostratigraphic scale has evolved since the mid-1800s and concerns the relative order of strata. Important events in its development were the realization by William Smith that in a horizontal sequence of sedimentary strata what is now an upper stratum was originally deposited on a lower one and the discovery by James Hutton that an unconformity (discontinuity) indicates a significant gap in time. Furthermore, the presence of fossils throughout Phanerozoic sediments has enabled paleontologists to construct a relative order of strata. As was explained earlier, at specific stratigraphic boundaries certain types of fossils either appear or disappear or both in some cases. Such biostratigraphic boundaries separate larger or smaller units of time that are defined as eons, eras, periods, epochs, and ages.

The chronometric scale is of more recent origin. It was made possible by the development of mass spectrometers during the 1920s and their use in geochronological laboratories for radiometric dating (see above). The chronometric scale is based on specific units of duration and on the numerical ages that are assigned to the aforementioned chronostratigraphic boundaries. The methods used entail the isotopic analyses of whole rocks and minerals of element pairs, such as potassium–argon, rubidium–strontium, uranium–lead, and samarium–neodymium. Another radiometric time scale has been developed from the study of the magnetization of basaltic lavas of the ocean floor. As such lavas were extruded from the mid-oceanic ridges, they were alternately magnetized parallel and opposite to the present magnetic field of the Earth and are thus referred to as normal and reversed. A magnetic-polarity time scale for the stratigraphy of normal and reversed magnetic stripes can be constructed back as far as the middle of the Jurassic Period, about 170 million years ago, which is the age of the oldest extant segment of ocean floor.

Time scale of magnetic-polarity reversals

During recent years, various chronostratigraphic and chronometric time scales with relatively small differences have been proposed. The 1983 scale prepared for the Decade of North American Geology (DNAG) has been selected for the present discussion because it takes into account many of the variations of other scales. As can be seen from Table 4, the DNAG geologic time scale gives the major chronostratigraphic boundaries and their assigned chronometric ages, as well as the time scale of magnetic polarity reversals.

Precambrian time

GENERAL CONSIDERATIONS

The Precambrian is defined as the period of time that extends from a little more than 3.9 billion years ago, which is the approximate age of the oldest known rocks, to the beginning of the Cambrian Period, roughly 540 million years ago. The Precambrian era thus represents more than 80 percent of the whole of geologic time. It has long been known that the Cambrian marks the earliest stage in the history of the Earth when many varied forms of life evolved and were preserved extensively as fossil remains in sedimentary rocks. It is not surprising that all life-forms were long assumed to have originated in the Cambrian, and therefore all earlier rocks with no obvious fossils were grouped together into one large era, the Precambrian. However, detailed mapping and examination of Precambrian rocks on most continents have since revealed that primitive life-forms already existed more than 3.5 billion years ago. The original terminology to distinguish Precambrian from all younger rocks, nevertheless, is still used for subdividing geologic time.

Major subdivisions. It is now internationally agreed that Precambrian time should be divided into the Archean and Proterozoic eons, with the time boundary between them at 2.5 billion years. The subdivision of these eons into early, middle, and late eras is not so widely agreed upon, but experts in the field have adopted a scheme according to which the relevant boundaries are at 3.4 billion and 3 billion and at 1.6 billion and 900 million years. These definitions are based on isotopic age determinations, and it is not possible to introduce smaller subdivisions. In the absence of fossils to permit the creation of small-scale subdivisions, relative chronologies of events have been produced for different regions based on such field relationships as unconformities and crosscutting dikes (tabular bodies of intrusive igneous rock that cut across original structures in the surrounding rock), combined with isotopic age determinations of specific rocks, as, for example, granites. This allows for some correlation between neighbouring regions.

Distinctive features. The Archean and Proterozoic are very different and must be considered separately. The Archean–Proterozoic boundary constitutes a major turning point in Earth history. Before that time the crust of the Earth was in the process of growing and so there were no large stable continents, whereas afterward, when such continents had emerged, orogenic belts were able to form marginally to and between continental blocks as they did during Phanerozoic times.

Archean orogenic belts

There are two types of Archean orogenic belts: (1) upper crustal greenstone–granite belts, rich in volcanic rocks which are probably primitive types of oceanic crust and island arcs that formed during the early rapid stage of crustal growth, and (2) granulite–gneiss belts that were recrystallized in the Archean mid-lower crust under metamorphic conditions associated with high-temperature granulite and amphibolite facies. Thus granulites, which typically contain the high-temperature mineral hypersthene, are a characteristic feature of many Precambrian orogenic belts that have been deeply eroded, as opposed to Phanerozoic orogenic belts, in which they are rare.

There are several other rock types that were developed primarily during the Precambrian and rarely later. This restriction is a result of the unique conditions that prevailed during Precambrian times. For example, the banded-iron formation mentioned above is a ferruginous sediment that was deposited on the margins of early iron-rich oceans. Anorthosite, which consists largely of plagioclase, forms large bodies in several Proterozoic belts. Komatiite is a magnesium-rich, high-temperature volcanic rock derived from a very hot mantle; it was extruded in abundance during the early Precambrian when the heat flow of the Earth was higher than it is today. Blueschist contains the blue mineral glaucophane; it forms in subduction zones under high pressures and low temperatures, and its rare occurrence in Precambrian rocks may indicate that temperatures in early subduction zones were too high for its formation.

The bulk of many of the world's valuable mineral deposits (for example, those of gold, nickel, chromite, copper, and iron) also formed during the Precambrian. These concentrations are a reflection of distinctive Precambrian sedimentary and magmatic rocks and their environments of formation.

PRECAMBRIAN ROCKS

General occurrence and distribution. Precambrian rocks as a whole occur in a wide variety of shapes and sizes. There are extensive Archean regions, up to a few thousands of kilometres across, that may contain either greenstone–granite belts or granulite–gneiss belts or both and that are variously designated in different parts of the world as cratons, shields, provinces, or blocks. Some examples are the North Atlantic craton that includes northwestern Scotland, central Greenland, and Labrador; the Kaapvaal and Zimbabwean cratons in southern Africa; the Dhārwār craton in India; the Aldan and Anabar shields in Siberia in Russia; the Baltic Shield that includes much of Sweden, Finland, and the Kola Peninsula of far northern Russia; the Superior and Slave provinces in Canada; and the Yilgarn and Pilbara blocks in Western Australia. There are linear belts, up to several thousand kilometres long, that are frequently though not exclusively of Proterozoic age, such as the Limpopo, Mozambique, and Damaran belts in Africa, the Labrador Trough in Canada, and the Eastern Ghāts belt in India. Also, small relict areas, only about

Notable Archean regions

Table 4: Geologic Time Scale

Cenozoic

age*	magnetic polarity: interval†	magnetic polarity: anomalies‡	magnetic polarity: polarity chron§	period	period	epoch	epoch		age	boundaries (Ma)
		1	C1	Quaternary		Holocene				0.01
						Pleistocene			Calabrian	1.6
		2 2A	C2 C2A	Tertiary	Neogene	Pliocene	L		Piacenzian	3.4
5		3 3A	C3				E		Zanclean	5.3
			C3A			Miocene	L		Messinian	6.5
10		4 4A 5	C4 C4A C5						Tortonian	11.2
15		5A	C5A				M		Serravallian	15.1
		5B	C5B						Langhian	16.6
20		5C 5D 5E 6	C5C C5D C5E C6				E		Burdigalian	21.8
		6A 6B	C6A C6B						Aquitanian	23.7
25		6C 7 7A 8 9	C6C C7 C7A C8 C9		Paleogene	Oligocene	L		Chattian	30.0
30 35		10 11 12 13	C10 C11 C12 C13				E		Rupelian	36.6
		15 16	C15 C16			Eocene	L		Priabonian	40.0
40		17 18	C17 C18						Bartonian	43.6
45 50		19 20 21	C19 C20 C21				M		Lutetian	52.0
55		22 23 24	C22 C23 C24				E		Ypresian	57.8
60		25	C25			Paleocene	L	Selandian	Thanetian	60.6
		26	C26						unnamed	63.6
65		27 28 29	C27 C28 C29				E		Danian	66.4

Mesozoic

age (Ma)	magnetic polarity: interval†	magnetic polarity: anomalies‡	magnetic polarity: polarity chron§	period	epoch		age	boundaries (Ma)	uncertainty (in millions of years)
		29 30	C29 C30					66.4	
70		31 32	C31 C32	Cretaceous	Late		Maastrichtian	74.5	4
80		33	C33				Campanian	84.0	
							Santonian	87.5	4.5
							Coniacian	88.5	
90							Turonian	91.0	2.5
							Cenomanian	97.5	2.5
100 110					Early		Albian	113	4
		M0					Aptian	119	9
120		M1				Neocomian	Barremian	124	9
130		M3 M5					Hauterivian	131	8
		M10 M12					Valanginian	138	5
140		M14 M16					Berriasian	144	5
		M18 M20		Jurassic	Late		Tithonian	152	12
150		M22					Kimmeridgian	156	6
160		M25 M29					Oxfordian	163	15
	rapid polarity changes				Middle		Callovian	169	15
170							Bathonian	176	34
180							Bajocian	183	34
							Aalenian	187	34
190					Early		Toarcian	193	28
							Pliensbachian	198	32
200							Sinemurian	204	18
							Hettangian	208	18
210 220				Triassic	Late		Norian	225	8
							Carnian	230	22
230					Middle		Ladinian	235	10
							Anisian	240	22
240					Early		Scythian	245	20

Paleozoic

age (Ma)	period	period	epoch	age		boundaries (Ma)	uncertainty (in millions of years)
						245	20
	Permian		Late	Tatarian		253	20
				Kazanian Ufimian		258	24
260			Early	Kungurian		263	22
				Artinskian		268	12
280				Sakmarian Asselian		286	12
	Carboniferous	Pennsylvanian	Late	Gzelian Kasimovian	S.	296	10
300				Moscovian	W.		
				Bashkirian		315	20
					N.	320	
320		Mississippian	Early	Serpukhovian		333	22
340				Visean		352	8
360				Tournaisian		360	10
	Devonian		Late	Famennian		367	12
				Frasnian		374	18
380			Middle	Givetian		380	18
				Eifelian		387	28
			Early	Emsian		394	22
400				Pragian		401	18
				Lochkovian		408	12
	Silurian		Late	Pridolian		414	12
420				Ludlovian		421	12
			Early	Wenlockian		428	8
440				Llandoverian		438	12
	Ordovician		Late	Ashgillian		448	12
460				Caradocian		458	16
			Middle	Llandeilan		468	16
480				Llanvirnian		478	16
			Early	Arenigian		488	20
500				Tremadocian		505	32
	Cambrian		Late	Trempealeauan Franconian Dresbachian		512	36
520			Middle			520	28
540			Early			540	

Precambrian

age (Ma)	eon	era	boundary ages (Ma)
			540
750	Proterozoic	Late	900
1,000 1,250 1,500		Middle	1,600
1,750 2,000 2,250		Early	2,500
2,750	Archean	Late	3,000
3,250		Middle	3,400
3,500 3,750		Early	3,800?

*Millions of years before the present (Ma). †Magnetic polarity interval where black indicates normal polarity and white signifies reversed polarity. ‡Marine magnetic anomalies—*i.e.*, those of the oceanic crust, with 1 starting at the mid-oceanic ridge where new crust is generated. §Term that describes the main subdivisions of time in continental sedimentary rocks recognized by their polarity and classified against the closest marine magnetic anomalies.

Published with permission of the Geological Society of America

a few hundred kilometres across, exist within or against Phanerozoic orogenic belts, as, for instance, the Lofoten islands of Norway, the Lewisian Complex in northwestern Scotland, and the Adirondack Mountains in the northeastern United States. Some extensive areas of Precambrian rocks are still overlain by a blanket of Phanerozoic sediments, as under the European and Russian platforms and under the central United States; these are mostly known from borehole samples (see Figure 15).

Adapted from Brian F Windley, *The Evolving Continents*, John Wiley and Sons Ltd

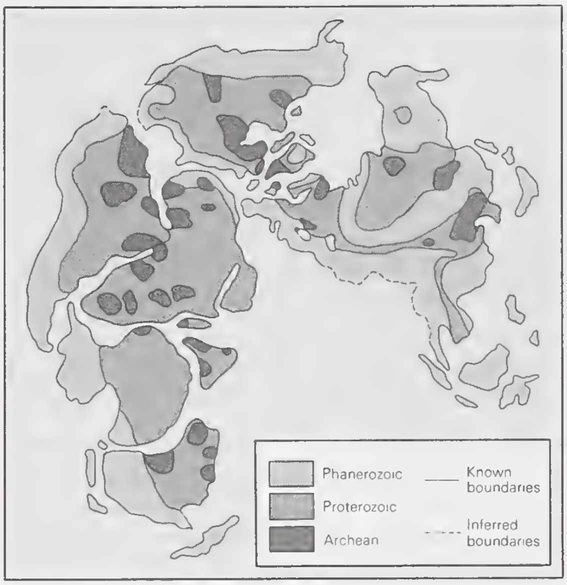


Figure 15: Archean regions within Proterozoic cratons surrounded by Phanerozoic mobile belts. This distribution is shown here on a Permian predrift map of the continents.

Archean rock types. Archean rocks occur in greenstone–granite belts that represent the upper crust, in granulite–gneiss belts that formed in the mid-lower crust, and in sedimentary basins, basic dikes, and layered complexes that were either deposited on or intruded into the first two types of belts.

Greenstone–granite belts. These belts occur on most continents. The largest extend several hundred kilometres in length and measure several hundred metres in width. They range from aggregates of several belts (as in the southern Superior province of Canada) to irregular, even triangular-shaped belts (such as Barberton in South Africa) to synclinal basins (as in the Indian Dhārwār craton). Today, many greenstone–granite belts are regarded as tectonic slices that have been thrust between or against older rocks, such as gneisses. The irregular and synclinal shapes are commonly caused by the diapiric intrusion of younger granites. Important occurrences are the Barberton belt in South Africa; the Sebakwian, Belingwean, and Bulawayan–Shamvaian belts of Zimbabwe; the Yellowknife belts in the Slave province of Canada; the Abitibi, Wawa, Wabigoon, and Quetico belts of the Superior province of Canada; the Dhārwār belts in India; and the Warrawoona belt in the Pilbara block and the Yilgarn belts in Australia.

Greenstone–granite belts developed at many different times throughout the long Archean Eon. In the Zimbabwean craton they formed over three successive periods: the Selukwe belt at about 3.75–3.8 billion years ago, the Belingwean belts at about 2.9 billion, and the Bulawayan–Shamvaian belts at 2.7–2.6 billion. The Barberton belt in the Kaapvaal craton and the Warrawoona belt in the Pilbara block are 3.5 billion years old. Globally the most important period of formation was from 2.7 to 2.6 billion years ago, especially in the Slave and Superior provinces of North America, the Yilgarn block in Australia, and the Dhārwār craton in India. Some of the better documented belts seem to have formed within about 50 million years. It is important to note that while the Bulawayan–Shamvaian belts were forming in the Zimbabwean craton, flat-lying sediments and volcanics were laid down in the Pongola rift and the Witwatersrand Basin not far to the north.

The greenstone sequence in many belts is divisible into a lower volcanic group and an upper sedimentary group. The volcanics are made up of ultramafic and basaltic lavas noted for magnesian komatiites that probably formed in the oceanic crust, overlain by basalts, andesites, and rhyolites whose chemical composition is much like that of modern island arcs. Especially important is the presence in the Barberton and Yellowknife belts of sheeted basic dike complexes cutting across gabbros and overlain by pillow-bearing basalts. The uppermost sediments are typically terrigenous shales, sandstones, quartzites, graywackes, and conglomerates. The overall stratigraphy suggests an evolution from extensive submarine eruptions of komatiite and basalt (ocean floor) to more localized stratovolcanoes (island arcs), which become increasingly emergent with intervening and overlying clastic basins. There are, however, regional differences in the volcanic and sedimentary makeup of some belts. The older belts in southern Africa and Australia have more komatiites and basalts, more shallow-water banded-iron formations, cherts, and evaporites, and fewer terrigenous sediments. On the other hand, the younger belts in North America have a higher proportion of andesites, rhyolites, and terrigenous and turbidite debris, but fewer shallow-water sediments. These differences reflect a change from the older oceanic-type volcanism (effusion of lava from submarine fissures) to the younger, more arc-type phenomenon (explosive eruption of pyroclastic materials and lava from steep volcanic cones), an increasing amount of trench turbidites and graywackes with time, and an increasing availability of continental crust with time as a source for terrigenous debris.

Regional differences in component rocks

Granitic rocks and gneisses occur within, adjacent to, and between many greenstone sequences. Some paragneisses, as in the Quetico belt in Canada, are derived from graywackes and were probably deposited in a trench or accretionary prism at the mouth of a subduction zone between the island arcs of the adjacent greenstone sequences. Many early granitic plutons were deformed and converted into orthogneiss. Late plutons commonly intruded the greenstones that were downfolded in synclines between them, or they intruded along the borders of the belts, deflecting them into irregular shapes.

The structure of many belts is complex. Their stratigraphic successions are upside down and deformed by thrusts and major horizontal folds (nappes) and have been subsequently refolded by upright anticlines and synclines. The result of this thrusting is that the stratigraphic successions have been repeated and thus may be up to 10–20 kilometres thick. Also, there may be thrusts along the base of the belts, as in the case of Barberton, showing that they have been transported from elsewhere. In other instances, the thrusts may occur along the borders of the belts, indicating that they have been forced against and over adjacent gneissic belts. The conclusion from structural studies is that many belts have undergone intense subhorizontal deformation during thrust transport and that subsequent compression, assisted by the diapiric rise of late granitic plutons, created synformal shapes and subvertical structures.

All the greenstone sequences have undergone recrystallization during metamorphism of greenschist facies at relatively low temperatures and pressures. In fact, the presence of the three green metamorphic minerals chlorite, hornblende, and epidote has given rise to the term greenstone for the recrystallized basaltic volcanics. Ultramafic rocks are commonly altered to talc schists and tremolite–actinolite schists, which may contain asbestos veins of economic value. There are some indications that several phases of metamorphism exist—namely, seafloor metamorphism associated with the action of hydrothermal brines perhaps at mid-oceanic ridges, syntectonic metamorphism related to thrust-nappe tectonics, and local thermal contact metamorphism caused by late intrusive granitic plutons.

One finds abundant mineralization in greenstone–granite belts. These belts constitute one of the world's principal depositories of gold, silver, chromium, nickel, copper, and zinc. In the past they were termed gold belts because the gold rushes of the 19th century took place, for example,

Mineralization in greenstone–granite belts

at Kalgoorlie in the Yilgarn belt of Western Australia, in the Barberton belt of South Africa, and at Val d'Or in the Abitibi belt of southern Canada. The mineral deposits occur in all the major rock groups: chromite, nickel, asbestos, magnesite, and talc in ultramafic lavas; gold, silver, copper, and zinc in basaltic to rhyolitic volcanics; iron ore, manganese, and barite in sediments; and lithium, tantalum, beryllium, tin, molybdenum, and bismuth in granites and associated pegmatites. Important occurrences are chromite at Selukwe in Zimbabwe, nickel at Kambalda in southwestern Australia, tantalum in Manitoba in Canada, and copper–zinc at Timmins and Noranda in the Canadian Abitibi belt.

Clearly there are different types of greenstone–granite belts. In order to understand their origin and mode of evolution, it is necessary to correlate them with comparable modern analogues. Some, like the Barberton and Yellowknife belts, consist of oceanic-type crust and have sheeted dike swarms that occur in many Mesozoic–Cenozoic ophiolites, such as Troodos in Cyprus. They are a hallmark of modern oceanic crust that formed at a mid-oceanic ridge. Also, like modern ophiolites, a few seem to have been obducted by thrusting onto continental crust. Many belts, like those in the Superior province of Canada, are very similar to modern island arcs. The Wawa belt, for example, has been shown to consist of an immature island arc built on oceanic crust and overlain by a more mature arc. The Abitibi belt began as an island arc that was rifted to form an intra-arc basin in which developed a second arc. Between the Wawa and Wabigoon island arcs lies the Quetico belt, consisting of metamorphosed turbidites and slices of volcanics that probably developed in an imbricated accretionary prism in an arc–trench system, as seen today in the Japanese arcs. The Pilbara belts are similar to modern active continental margins, and they have been interthrust with older continental orthogneisses to form a very thick crustal pile that was intruded by diapiric crustal-melt granites. This scenario is quite comparable to that of a Himalayan type of orogenic belt formed by collisional tectonics. In conclusion, most greenstone–granite belts are today regarded by geologists as different parts of interthrust oceanic trench–island arc systems that collided with continental gneissic blocks.

High-grade metamorphism

Granulite–gneiss belts. The granulites, gneisses, and associated rocks in these belts were metamorphosed to a high grade in deep levels of the Archean crust; metamorphism occurred at a temperature of 750° to 980° C and at a depth of about 15–30 kilometres. These belts, therefore, represent sections of the continents that have been highly uplifted, with the result that the upper crust made up of volcanics, sediments, and granites has been eroded. Accordingly, the granulite–gneiss belts are very different from the greenstone–granite belts. They occur in a variety of environments. These may be extensive regions, such as the North Atlantic craton, which measures 1,000 by 2,000 kilometres across and includes, in its pre-Atlantic fit, the Scourian complex of northwestern Scotland, the central part of Greenland, and the coast of Labrador; the Aldan and Ukrainian shields of continental Europe; eastern Hopeh (Hebei) and Liaoning provinces of northeastern China; large parts of the Superior province of Canada and the Yilgarn block in Australia; and the Limpopo belt in southern Africa. Or they may be small areas, such as the Ancient Gneiss Complex of Swaziland; the Minnesota River valley and the Beartooth Mountains of the United States; the Peninsular gneisses and Sargur supracrustals of southern India; the English River gneisses of Ontario in Canada that form a narrow strip between greenstone–granite belts; the Sand River gneisses that occupy a small area between greenstone–granite belts in Zimbabwe; and the Napier complex in Enderby Land in Antarctica. Granulite–gneiss belts are commonly surrounded by younger, mostly Proterozoic belts that contain remobilized relicts of the Archean rocks, and the granulites and gneisses must underlie many Archean greenstone–granite belts and blankets of Phanerozoic sediment. In light of this, the granulite–gneiss belts may be regarded as minimally exposed and preserved sections of a widespread continental basement.

Isotopic age determinations from the granulite–gneiss belts record an evolution from about 4.3 to 2.5 billion years—more than a third of geologic time. Most important are the few but well-determined detrital zircons at Mount Narryer and Jack Hills in Western Australia that are more than 4 billion years old (see above). Several regions have a history that began in the period dating from 3.9 to 3.6 billion years ago—*e.g.*, West Greenland, Labrador, the Limpopo belt, Enderby Land, and the Aldan Shield. Most regions of the world went through a major tectonic event that may have involved intrusion, metamorphism, and deformation in the period between 3.1 and 2.8 billion years ago; and some of these regions, like the Scourian in northwestern Scotland, show no evidence of any older crustal growth. The best-documented region is West Greenland, which has a long and complicated history from 3.8 to 2.5 billion years ago.

Most common rock type

Orthogneisses of deformed and recrystallized tonalite and granite constitute the most common rock type. The geochemical signature of these rocks closely resembles that of modern equivalents that occur in granitic batholiths in the Andes. Where such rocks have been metamorphosed under conditions associated with amphibolite facies, they contain hornblende or biotite or a combination of the two. However, where they have been subjected to conditions of higher temperature associated with the granulite facies, the rocks contain pyroxene and hypersthene and so can be called granulites.

The granulites and gneisses enclose a wide variety of other minor rock types in layers and lenses. These types include schists and paragneisses, which were originally deposited on the Earth's surface as shales and which now contain high-temperature metamorphic minerals, such as biotite, garnet, cordierite, staurolite, sillimanite, or kyanite. There also are quartzites, which were once sandstones; marbles, which were either limestones or dolomites; and banded-iron formations, which were deposited as ferruginous sediments. Commonly intercalated with these metasediments are amphibolites, which locally contain relict pillows, demonstrating that they are derived from basaltic lavas deposited underwater. These amphibolites have a trace element chemistry quite similar to that of modern seafloor basalts. The amphibolites are often accompanied by chromite-layered anorthosite, gabbro, and ultramafic rocks, such as peridotite and dunite. All these rocks occur in layered igneous complexes, which in their well-preserved state may be up to 1 kilometre thick and 100 kilometres long. Such complexes occur at Fiskenaesset in West Greenland and in the Limpopo belt in southern Africa, as well as in southern India. These complexes may have formed at a mid-oceanic ridge in a magma chamber that also fed the basaltic lavas. In many cases, the complexes, the basaltic amphibolites, and the sediments were extensively intruded by the tonalites and granites that were later deformed and recrystallized, with the result that all these rocks may now occur as metre-sized lenses in the orthogneisses and granulites.

The structure of the granulite–gneiss belts is extremely complex, since the constituent rocks have been highly deformed several times. In all likelihood, it seems that the basalts and layered complexes from the oceanic crust were interthrust with shallow-water limestones, sandstones, and shales, with tonalites and granites from Andean-type batholiths, and with older basement rocks from a continental margin. All these rocks, which are now mutually conformable, were folded in horizontal nappes and then refolded. The picture that emerges is one of a very mobile Earth, where no rocks remained long after their formation before they were compressed and thrust against other rocks.

The mid-lower crust is relatively barren of ore deposits, as compared to the upper crust with its sizable concentrations of greenstones and granites, and so little mineralization is found in the granulite–gneiss belts. The few exceptions include a nickel–copper sulfide deposit at Selebi–Pikwe in the Limpopo belt in Botswana that is economic to mine and banded-iron formations in gneisses in the eastern Hopeh and Liaoning provinces of northwestern China that form the foundation of a major steel industry. There are uneconomic chromitite seams in anorthosites in West Greenland, southern India, and the Limpopo belt, a banded-iron formation in a 3.8-billion-year-old sedimentary–volcanic belt at Isua in West Greenland, and minor tungsten mineralization in amphibolites in West Greenland.

Correlation

It is impossible to correlate the rocks in different granulite–gneiss belts. One granitic gneiss is essentially the same as another, but it may be of vastly different age. There is a marked similarity in the anorthosites in various belts throughout the world, and their similar relationship with the gneisses suggests that the belts have undergone comparable stages of evolution, although each has its own distinctive features. Little correlation can be made with rocks of Mesozoic–Cenozoic age, because few modern orogenic belts have been eroded sufficiently to expose their mid-lower crust. The lack of modern analogues for comparison makes it particularly difficult to interpret the mode of origin and evolution of the Archean granulite–gneiss belts.

Sedimentary basins, basic dikes, and layered complexes. During the Late Archean (3 to 2.5 billion years ago), relatively stable, post-orogenic conditions developed locally in the upper crust, especially in southern Africa, where the development of greenstone–granite and granulite–gneiss belts was completed much earlier than in other parts of the world. The final chapters of Archean crustal evolution can be followed by considering specific key sedimentary basins, basic (basaltic) dikes, and layered complexes.

Pongola Rift

Along the border of Swaziland and South Africa is the Pongola Rift, which is the oldest such continental trough in the world; it is 2.95 billion years old, having formed only 50 million years after the thrusting of adjacent greenstone–granite belts. If there were earlier rifts, they have not survived, or more likely this was the first time in Earth history that the upper crust was sufficiently stable and rigid for a rift to form. It is 30 kilometres wide and 130 kilometres long, and within it is an 11-kilometre-thick sequence of lavas and sediments. It seems most likely that the rift developed as the result of the collapse of an overthickened crust following the long period of Archean crustal growth and thrusting in the Kaapvaal craton.

The 200-by-350-kilometre Witwatersrand Basin contains an 11-kilometre-thick sequence of lavas and sediments that are 2.8 billion years old. The basin is famous for its very large deposits of gold and uranium that occur as detrital minerals in conglomerates. These minerals were derived by erosion of the surrounding greenstone–granite belts and transported by rivers into the shoreline of the basin. In all probability, the gold originally came from the komatiitic and basaltic lavas in the early Archean oceanic crust.

Great Dyke of Zimbabwe

The Great Dyke, thought to be more than 2.5 billion years old, transects the entire Zimbabwe craton. It is 480 kilometres long and 8 kilometres wide and consists of layered ultrabasic rocks—gabbros and norites. The ultrabasic rocks have several layers of chromite and an extensive platinum-bearing layer that form economic deposits. The Great Dyke represents a rift that has been filled in with magma.

The Stillwater Complex is a famous, 2.7-billion-year-old, layered ultrabasic-basic intrusion in the Beartooth Mountains of Montana, U.S. It is 48 kilometres long and has a stratigraphic thickness of 6 kilometres. It was intruded as a subhorizontal body of magma that underwent crystal settling to form the layered structure. It is notable for a three-metre-thick layer enriched in platinum minerals, which forms a major economic deposit.

The basins, dikes, and complexes described above cannot be mutually correlated. They most resemble equivalent structures that formed at the end of plate-tectonic cycles in the Phanerozoic. They represent the culmination of Archean crustal growth.

Proterozoic rock types. What happened geologically at the time of the Archean–Proterozoic boundary 2.5 billion years ago is uncertain. It seems to have been a period of little tectonic activity, and so it is possible that the earlier intensive Archean crustal growth had caused the amalgamation of continental fragments into a supercontinent, perhaps similar to Pangaea in the Permo–Triassic. The fragmentation of this supercontinent and the formation of new oceans gave rise to many continental margins on which a variety of distinctive sediments were deposited. Much evidence suggests that in the period from 2.5 billion to 570 million years ago Proterozoic oceans were formed and destroyed by plate-tectonic processes and that most Proterozoic orogenic belts arose by collisional tectonics. Sedimentary, igneous, and metamorphic rocks that formed in this period are widespread throughout the world. There are many swarms of basic dikes, important sedimentary rifts, basins, and layered igneous complexes, as well as many orogenic belts. The rocks commonly occur in orogenic belts that wrap around the borders of Archean cratons. The characteristic types of Proterozoic rocks are considered below, as are classic examples of their occurrence in orogenic belts. With a few exceptions the following types of rocks were formed during the Early, Middle, and Late Proterozoic, indicating that similar conditions and environments existed throughout this long period of time.

Basic dikes. The continents were sufficiently stable and rigid during the Proterozoic for an extremely large number of basic dikes to be intruded into parallel, extensional fractures in major swarms. Individual dikes measure up to several hundred metres in width and length, and there may be hundreds or even thousands of dikes in a swarm, some having transcontinental dimensions. For example, the 1.2-billion-year-old Mackenzie swarm is more than 500 kilometres wide and 3,000 kilometres long and extends in a northwesterly direction across the whole of Canada from the Arctic to the Great Lakes. The 1.95-billion-year-old Kangamiut swarm in West Greenland is only about 250 kilometres long but is one of the world's densest continental dike swarms. Many of the major dike swarms were intruded on the continental margins of Proterozoic oceans in a manner similar to the dikes that border the present-day Atlantic Ocean.

Layered igneous intrusions. There are several very important layered, mafic to ultramafic intrusions of Proterozoic age that were formed by the accumulation of crystals in large magma chambers. The well-known ones are several tens or even hundreds of kilometres across, have a dikelike or sheetlike (stratiform) shape, and contain major economic mineral deposits. The largest and most famous is the Bushveld Complex in South Africa, which is 9 kilometres thick and covers an area of 66,000 square kilometres. It was intruded nearly 2.1 billion years ago and is the largest repository of magmatic ore deposits in the world. The Bushveld Complex consists of stratiform layers of dunite, norite, anorthosite, and ferrodiorite and contains deposits of chromite, iron, titanium, vanadium, nickel, and, most important of all, platinum. The Sudbury Complex in southern Canada, which is about 1.9 billion years old, is a basin-shaped body that extends up to 60 kilometres across. It consists mostly of layered norite and has deposits of copper, nickel, cobalt, gold, and platinum. It is noted for its high-pressure structures and other manifestations of shock metamorphism, which suggest that the intrusion was produced by an enormous meteorite impact (see CONTINENTAL LANDFORMS: *Formation of impact craters*).

World's largest repository of magmatic ore deposits

Shelf-type sediments. Quartzites, dolomites, shales, and banded-iron formations make up sequences that reach up to 10 kilometres in thickness and that amount to more than 60 percent of Proterozoic sediments. Minor sediments include sandstones, conglomerates, red beds, evaporites, and cherts. The quartzites typically have cross-bedding and ripple marks, which are indicative of tidal action, and the dolomites often contain stromatolites similar to those that grow today in intertidal waters. Also present in the dolomites are phosphorites that are similar to those deposited on shallow continental margins against areas of oceanic upwelling during the Phanerozoic. Several early-middle Proterozoic examples of such dolomites have been found in Finland and northern Australia, as well as in the Marquette Range of Michigan in the United States, in the Aravalli Range of Rājasthān in northwestern India, and at Hamersley and Broken Hill in Australia. Still another constituent of these dolomites is evaporite, which contains casts and relicts of halite, gypsum, and anhydrite and which occurs, for example, at Mount Isa in Australia (1.6 billion years old) and in the Belcher Group in Canada (1.8 billion years old). These evaporites were deposited by brines in very shallow pools, like those encountered today in the Persian Gulf.

Ophiolites. Phanerozoic ophiolites are considered to be fragments of ocean floor that have been trapped between island arcs and continental plates that collided or that have been thrust onto the shelf sediments of continental margins. They consist of a downward sequence of oceanic sediments such as cherts, pillow-bearing basalts, sheeted (100-percent) basic dikes, gabbros, and certain ultramafic rocks (*e.g.*, serpentinized harzburgite and lherzolite). Comparable ophiolites occur in several Proterozoic orogenic belts and provide strong evidence of the existence of oceanic plates like those of today. The oldest is an ophiolite in the Cape Smith belt on the south side of Hudson Bay in Canada whose age has been firmly established at 1.999 billion years. There is a 1.8-billion-year-old ophiolite in the Svecofennian belt of southern Finland, but most Proterozoic ophiolites are 1 billion to 570 million years old and occur in the Pan-African belts of Saudi Arabia, Egypt, and The Sudan, where they occur in sutures between a variety of island arcs.

Greenstones and granites. Greenstone–granite belts, like those of the Archean, continued to form in the Proterozoic albeit in greatly reduced amounts. They are characterized by abundant volcanic rocks that include pillowed subaqueous basalt flows and subaerial and subaqueous volcaniclastic rocks. Magnesian komatiites are for the most part absent, however. Intrusive plutons are typically made of granodiorite. Examples occur at Flin Flon in central Canada and in the Birrimian Group in West Africa. Generally such rocks resemble those in modern island arcs and back-arc basins.

Intrusive plutons of granodiorite

Granulites and gneisses. These highly deformed and metamorphosed rocks are similar to those of the Archean and occur in many Proterozoic orogenic belts, such as the Grenville in Canada, the Pan-African Mozambique belt in eastern Africa, the Musgrave and Arunta ranges in Australia, and in Lapland in the northern Baltic Shield. They were brought up from the mid-lower crust on major thrusts as a result of continental collisions.

Orogenic belts. Some of the classic Proterozoic orogenic belts of the world are considered in this section.

One such belt is the Wopmay Orogen, situated in the Arctic in the northwestern part of the Canadian Shield. (It is beautifully exposed and has been well described by Paul F. Hoffman and his colleagues at the Geological Survey of Canada.) It formed within a relatively short time between 1.9 and 1.8 billion years ago and provides convincing evidence of tectonic activity of a modern form in the early Proterozoic. On the eastern continental margin occur red beds (sandstones), which pass oceanward and westward into stromatolite-rich dolomites deposited on the continental shelf to a thickness of four kilometres; these dolomites pass into submarine turbidite fans that were deposited on the continental rise. An island arc and a continental margin are found to the west. The history of the Wopmay Orogen can be best interpreted in terms of subduction of oceanic crust and collision tectonics.

The Svecofennian Orogen of the Baltic Shield extends in a southeasterly direction from northern Sweden through southern Finland to the adjoining part of western Russia. It formed in the period from 1.9 to 1.7 billion years ago. A major lineament across southern Finland consists of the suture zone on which occur ophiolite complexes representing the remains of oceanic crust. At Outokumpu one encounters copper mineralization in these oceanic crust rocks similar to that in the Cretaceous ophiolite at Troodos in Cyprus. On the northern side of the suture is a shelf-type sequence of sediments and on the southern side a volcanic-plutonic arc. To the south of this arc lies a broad zone with thrusted gneisses intruded by tin-bearing crustal-melt granites, called rapakivi granites after their coarse, zoned feldspar megacrysts (*i.e.*, crystals that are significantly larger than the surrounding fine-grained matrix). The rocks in this zone are predictably equivalent to those that occur today under the Tibetan Plateau at a depth of about 20 kilometres.

Grenville Orogen

The Grenville Orogen is a deeply eroded and highly uplifted orogenic belt that extends from Labrador in northeastern Canada to the Adirondack Mountains and southwestward under the coastal plain of the eastern United States. It developed from about 1.5 to 1 billion years ago. Apart from an island arc situated today in Ontario, most of the Grenville Orogen consists of highly metamorphosed and deformed gneisses and granulites, which have been brought to the present surface on major thrusts from the mid-lower crust. A result of the terminal continental collision that occurred at about 1.1 billion years ago was the formation of the Midcontinent (or Keweenawan) rift system that extends southward for more than 2,000 kilometres from Lake Superior.

A type of crustal growth—one very different from that described above—took place in what is now Saudi Arabia, Egypt, and The Sudan in the period from 1.1 billion to 500 million years ago. This entire shield, called the Arabian–Nubian Shield, is dominated by volcanic lavas, tuffs (consolidated rocks consisting of pyroclastic fragments and ash), and granitic plutons that formed in a variety of island arcs separated by several sutures along which occur many ophiolite complexes. Some of the ophiolites contain a complete stratigraphy that is widely accepted as a section through the oceanic upper mantle and crust. The final collision of the arcs was associated with widespread thrusting and followed by the intrusion of granitic plutons containing tungsten, tin, uranium, and niobium ore deposits. The island arcs grew from the subduction of oceanic crust in a manner quite comparable to that taking place today throughout Indonesia.

The Mozambique belt is one of the many Pan-African orogenic belts that formed in the period between 1 billion and 500 million years ago. It extends along the eastern border of Africa from Ethiopia to Kenya and Tanzania. It consists largely of highly metamorphosed, mid-crustal gneisses within which are a few peridotite bodies that may be relicts of ophiolites. The structure of the Mozambique belt is dominated by eastward-dipping thrusts very similar to the thrusts on the southern side of the Himalayas that resulted from the collision of India with Tibet during the Tertiary Period hundreds of millions of years later.

Riphean and Sinian basins

During the Middle and Late Proterozoic, thick sequences of sediment were deposited in many basins throughout Asia. The Riphean sequence spans the period from 1.7 billion to 900 million years ago and occurs in what was formerly the Soviet Union. The Sinian in China extends from 850 to 540 million years ago (the end of the Precambrian) and roughly approximates the Vendian in Russia and certain other one-time Soviet states. The sediments are terrigenous debris characterized by conglomerates, sandstone, siltstone, and shale, some of which are oxidized red beds, along with stromatolite-rich dolomite. Total thicknesses reach over 10 kilometres. The terrigenous sediments were derived from the erosion of Proterozoic orogenic belts.

Correlation. The fact that Phanerozoic sediments have been so successfully subdivided and correlated is attributable to the presence of abundant fossil remains of life-forms that evolved and underwent changes over time. The fact that Precambrian sediments lack such fossils prevents any comparable correlations. There are, however, stromatolites in Precambrian sediments ranging in age from about 3.5 billion to 540 million years, and they reached their peak of development in the Proterozoic. Stromatolites underwent sufficient evolutionary changes that Russian biostratigraphers were able to use them to subdivide the Riphean into four main zones throughout widely separated areas of former Soviet territory. Similar stromatolite-based stratigraphic divisions have been recognized in the Norwegian islands of Spitsbergen, China, and Australia. This stromatolite biostratigraphy still has relatively limited application, however. As a consequence, it is the chronometric time scale that is used to subdivide Precambrian time and to correlate rocks from region to region and from continent to continent.

The rocks within Proterozoic orogenic belts are invariably too deformed to allow correlation of units between different belts. Nonetheless, the techniques of geochronology have improved considerably in recent years, with the result that rocks of approximately similar age on different continents can be mutually compared and regarded as equivalent. Archean rocks have in general been far too highly deformed and metamorphosed to be correlated to any significant degree.

PRECAMBRIAN ENVIRONMENT

In this section the types of environment that may have existed during Precambrian time are considered. Several rock types, notably banded-iron formations, paleosols, and red beds, are very useful for deriving information about the conditions of the atmosphere, and tillites (indurated sedimentary rocks formed by the lithification of glacial till; see below) reveal what the climatic patterns were like during Precambrian glaciations.

Continental drift and sediment deposition

Paleogeography. One of the most important factors controlling the nature of sediments deposited today is continental drift. This follows from the fact that the continents are distributed at different latitudes, and latitudinal position affects the temperature of oceanic waters along continental margins; in short, sedimentary deposition is climatically sensitive. At present, most carbonates and oxidized red soils are being deposited within 30 degrees of the equator, phosphorites within 45 degrees of it, and evaporites within 50 degrees. Most fossil carbonates, evaporites, phosphorites, and red beds of Phanerozoic age dating back to the Cambrian have a similar bimodal distribution with respect to their paleoequators. If the uniformitarian principle that the present is the key to the past is valid, then in the Precambrian such sediments would have likewise been controlled by the movement and geographic position of the continents. Thus it can be inferred that the stromatolite-bearing dolomites of the Riphean in the former Soviet Union were deposited in warm tropical waters. Even the 3.5-billion-year-old, extensive evaporites in the Pilbara region of northwestern Australia could not have been formed close to their paleopole. Today, phosphate sediments are deposited primarily along the western side of continents, where they receive upwelling, nutrient-rich currents as they move toward the equator. The major phosphorite deposits in the Proterozoic Aravalli belt of Rājasthān in northwestern India are associated with stromatolite-rich dolomites and were most likely deposited within the tropics on the western side of a continental mass.

Significant geologic events. Outlined below are some of the main geologic events that occurred throughout the long history of the Precambrian and that reveal something about prevailing conditions and environments.

Oldest minerals and rocks. As was previously noted, the oldest minerals on Earth, the zircons from western Australia, crystallized 4.276 billion years ago. The most significant thing about them is that the environment in which they formed is totally unknown. The rocks from which they came may have been destroyed by some kind of tectonic process or by a meteorite impact. On the other hand, the rocks may still exist on the Earth's surface but simply have not been found. Perhaps their very absence is indicative of something important about early terrestrial processes. Comparisons with the Moon indicate that the Earth must have been subjected to an enormous number of meteorite impacts about 4 billion years ago, but there is no geologic evidence of such events.

The oldest known rocks on Earth are found near Canada's Great Slave Lake; their age has been established radiometrically at 3.96 billion years. These rocks are of a granitic variety and are thought to have evolved from older basaltic crustal material that was melted and remelted by tectonic processes.

Archean events. During the first third of geologic history, until about 2.5 billion years ago, the Earth developed in a broadly similar manner. Greenstone–granite belts formed in the upper Archean crust and granulite–gneiss belts in the mid-lower crust. This was a time when the overall rate of heat production by the breakdown of radioactive isotopes was several times greater than it is today. This condition was manifested by very rapid tectonic processes, probably by some sort of primitive plate tectonics. Most of the heat that escapes from the Earth today does so at the mid-oceanic ridges, and it probably did likewise during the Archean but in much larger amounts. To permit this release of heat, the mid-oceanic ridges of the Archean were more abundant and longer and opened faster than those in the modern oceans. Although the amount of newly generated crust was probably enormous, a large part of this material was inevitably destroyed by equally rapid plate subduction processes. The main results of this early growth that can still be seen today are the many island arcs in greenstone–granite belts and the voluminous Andean-type tonalites that were deformed to orthogneiss in granulite–gneiss belts. Although most of the Archean oceanic crust was subducted, a few ophiolitic-type complexes have been preserved in greenstone–granite belts.

Rapid tectonic processes

The Late Archean was an important interval of time because it marks the beginning of the major changeover from Archean to Proterozoic types of crustal growth. Significant events of this time were the formation of the first major rifts (such as the Pongola), the intrusion of the first major basic dikes (*e.g.,* the Great Dyke) and of the first large stratiform layered igneous complexes (*e.g.,* the Stillwater), and the formation of the first large sedimentary basins (as, for example, the Witwatersrand). All of these structures indicate that the continental crust had for the first time reached a mature stage with considerable stability and rigidity. The Late Archean represents the culmination that followed the rapid tectonic processes of the Early and Middle Archean. Because crustal growth was diachronous (*i.e.,* cut across time planes) throughout the world, similar structures can be found in the Early Proterozoic.

The Archean–Proterozoic boundary. There is no record of tectonic activity of any sort at the time corresponding to the Archean–Proterozoic boundary—about 2.5 billion years ago. This probably means that a supercontinent was created by the amalgamation of innumerable smaller continental blocks and island arcs. Accordingly, this was a period of tectonic stability that may have been comparable to the Permo–Triassic when the supercontinent of Pangaea existed. The main geologic events would have been the intrusion of basic dikes and the formation of sedimentary basins, like the Huronian on the U.S.–Canadian border, into which large volumes of clastic sediment were deposited. Such sediments would have been derived by erosion of high plateaus and mountains that are characteristic of a large continental mass.

Proterozoic developments. During the Early Proterozoic large amounts of quartzite, carbonate, and shale were deposited on the shelves and margins of many continental blocks. This would be consistent with the breakup of a supercontinent into several or many smaller continents with long continental margins. Examples of shelf sequences of this kind are found along the margins of orogenic belts, such as the Wopmay bordering the Slave province and the Labrador Trough bordering the Superior province in Canada, and the Svecofennian in Finland.

Development of long, linear mountain belts

The fact that stable continental blocks existed by the Early Proterozoic meant that orogenic belts were able to develop against them by some form of collision tectonics. This was the first time that long, linear orogenic belts could form by "modern-style" plate-tectonic processes that involved seafloor spreading, ophiolite obduction, subduction that created island arcs and Andean-type granitic batholiths in active continental margins, and the collision of arcs and continents that gave rise to sutures with ophiolites and to Himalayan-type thrust belts with abundant crustal-melt granites. These were key events in the evolution of the continents, and such processes have continued throughout Earth history.

During the Late Proterozoic some orogenic belts continued to develop, as in the case of the Pan-African belts of Saudi Arabia and East Africa. The intense crustal growth and the many orogenic belts that formed throughout the Proterozoic, however, began to create large continental blocks, which amalgamated to form a new supercontinent by the end of the Precambrian. Therefore, in the Late

Proterozoic many sedimentary basins that were infilled with conglomerates and sandstones—the aforementioned Riphean and Sinian, for example—were able to form on extensive cratons of continental crust.

Climatic conditions. During the long course of Precambrian time the climatic conditions of the Earth must have changed considerably. Evidence of this can be seen in the sedimentary record, which suggests that the composition of the atmosphere and oceans changed appreciably over time. More importantly, however, the presence of tillites indicates that extensive glaciations occurred several times during the Precambrian. The tillites provide evidence of glacial conditions, although not necessarily at high latitudes. In general, they are complementary to the carbonates, evaporites, and red beds that are climatically sensitive and restricted to low latitudes.

Evidence of Precambrian glaciation

The oldest extensive glaciation occurred 2.3 billion years ago during the Early Proterozoic. It can be recognized from the rocks and structures that the glaciers and ice sheets left behind on several continents. The most extensive occurrences are found in North America in a belt nearly 3,000 kilometres long extending from Chibougamau in Quebec through Ontario to Michigan and southwestward to the Medicine Bow Mountains of Wyoming. This probably represents the area of the original ice sheet. Most details are known from the Gowganda Formation in Ontario, which contains glacial deposits that are up to 200 metres thick and that occupy an area of about 20,000 square kilometres. Evidence that these rocks were of glacial origin has been obtained by comparing them with the rocks left behind by the Quaternary ice sheets and with the deposits associated with modern glaciers. The Early Proterozoic examples have the following features: The main glacial sediment is a tillite. This lithified till contains abundant pebbles and fragments of up to boulder size of various rocks distributed randomly in a fine-grained silty matrix. The surfaces of some pebbles have parallel scratches caused by having been rubbed against harder pebbles during ice transport. Locally, the basement rocks below the tillite also have been scratched, or striated, by the movement of the overlying boulder-strewn ice. Another type of glacial deposit is a laminated varved sediment composed of alternating millimetre-to-centimetre-thick layers of silt and clay, which closely resemble the layered varves that are laid down in modern glacial lakes at the front of retreating glaciers or ice sheets, each layer defining an annual accumulation of sediment. Within the Gowganda varved sediment are dropstones, which are fragments of rock that have dropped from an overlying floating ice sheet and that have sunk into and depressed the varved layers beneath them. When all these features are found together, they provide good evidence of an ancient glaciation. Similar, roughly contemporaneous glacial deposits can be found in the Transvaal and Cape Province in South Africa, where they reach only 30 metres in thickness but extend over an area of 20,000 square kilometres. Such deposits also are encountered in the Hamersley Basin of Western Australia, in east-central Finland and the adjoining part of northwestern Russia, near Lake Baikal in Siberia, and in central India. These occurrences suggest that there was one or more extensive glaciation during the Early Proterozoic.

The largest glaciation in the history of the Earth occurred during the Late Proterozoic in the period between 1 billion and 600 million years ago. It left its mark almost everywhere. The principal occurrences of the glacial deposits are in Europe (Scotland, Ireland, Sweden, Norway, France, the Czech Republic, and Slovakia), the Western Cordillera (Yukon Territory, Can., to California, U.S.) and the Appalachians of the United States, East Greenland, Brazil, much of Africa (Congo, Angola, Namibia, Zambia, Zaire, and South Africa), and much of Russia, China, and Australia. One of the best-described occurrences is in the Flinders Range of South Australia, where there is a 4-kilometre-thick sequence of tillites and varved sediments occupying an area of 400-by-500 kilometres. Detailed stratigraphy and isotopic dating show that three glaciations took place at 850–800, 750, and 720–670 million years ago. The Port Askaig tillite on the island of Islay off northwestern Scotland is only 750 metres thick, but it records 17 ice advances and retreats and 27 periglacial periods, which are indicated by infilled polygons that formed under ice-free permafrost conditions. There are two major tillites in central Africa and Namibia (910–870 and 720–700 million years old, respectively) and two other such consolidated tills in East Greenland. What is the explanation for all these occurrences? It is interesting that some paleomagnetic studies have shown that the tillites in Scotland, Norway, Greenland, central Africa, North America, and South Australia were deposited in low or near-equatorial paleolatitudes. Such conclusions are, however, controversial, because it has also been suggested that the poles may have migrated across the globe, leaving a record of glaciations in high and low latitudes. There is the possibility that floating ice sheets could have traveled to low latitudes, depositing glacial sediments and dropstones below them. Whatever the answer, the existence of such vast quantities of tillites and of such extensive glaciations is intriguing and enigmatic. There is at present no broad agreement that resolves this Late Proterozoic phenomenon.

PRECAMBRIAN LIFE

Precambrian rocks were long ago defined to predate the Cambrian and therefore to predate all life, although the term Proterozoic was later coined from the Greek for "early life." It is now known that Precambrian rocks do in fact contain the evidence of the very beginnings of life on Earth (and thus the record of its evolution for more than 3 billion years), of the explosion of life-forms without skeletons before the Cambrian, and even of the development of sexual reproduction on Earth.

The first evidence of terrestrial life is found in the Early Archean sedimentary rocks of the greenstone–granite belts of Barberton in South Africa and of Warrawoona in the Pilbara block of Western Australia, which are both about 3.5 billion years old. There are two types of these early, simple, biological structures: microfossils and stromatolites.

Microfossils and stromatolites

The microfossils occur in cherts and shales and are of two varieties. One type consists of spherical carbonaceous aggregates, or spheroids, which may measure as much as 20 millimetres in diameter. These resemble algae and cysts of flagellates and are widely regarded as biogenic. The other variety of microfossils consists of carbonaceous filamentous threads, which are curving, hollow tubes up to 150 micrometres (0.006 inch) long. These tubes are most likely the fossil remains of filamentous organisms, and hundreds of them can be found in some rock layers. The 2.8-billion-year-old goldreefs (conglomerate beds with rich gold deposits) of the Witwatersrand Basin contain carbonaceous columnar microfossils up to seven millimetres long that resemble modern algae, fungi, and lichens. They probably extracted gold from the environment in much the way that modern fungi and lichens do.

Stromatolites are, as previously explained, stratiform, domal, or columnar structures made of sheetlike mats precipitated by communities of microorganisms, particularly filamentous blue-green algae. The Early Archean examples form domes as tall as about 10 centimetres (see Figure 16). Stromatolites occur in many of the world's greenstone–granite belts. In the 2.7-billion-year-old Steep Rock Lake belt in Ontario, Can., they reach three metres in height and diameter. Stromatolites continued to form all the way through the geologic record and today grow in warm intertidal waters, for example, at Shark Bay in Western Australia. They provide indisputable evidence that by 3.5 billion years ago life had begun on Earth by algal photosynthesis in complex, integrated biological communities.

These Archean organisms were prokaryotes that were incapable of cell division. They were relatively resistant to ultraviolet radiation and were able to survive during the early history of the Earth when the atmosphere lacked an ozone layer to block out such radiant energy. The prokaryotes were predominant until about 1.4 billion years ago, when they were overtaken by the eukaryotes. The latter make use of oxygen in metabolism and for growth and thus developed profusely in the increasingly oxygenic atmosphere of the Middle Proterozoic. The eukaryotes were

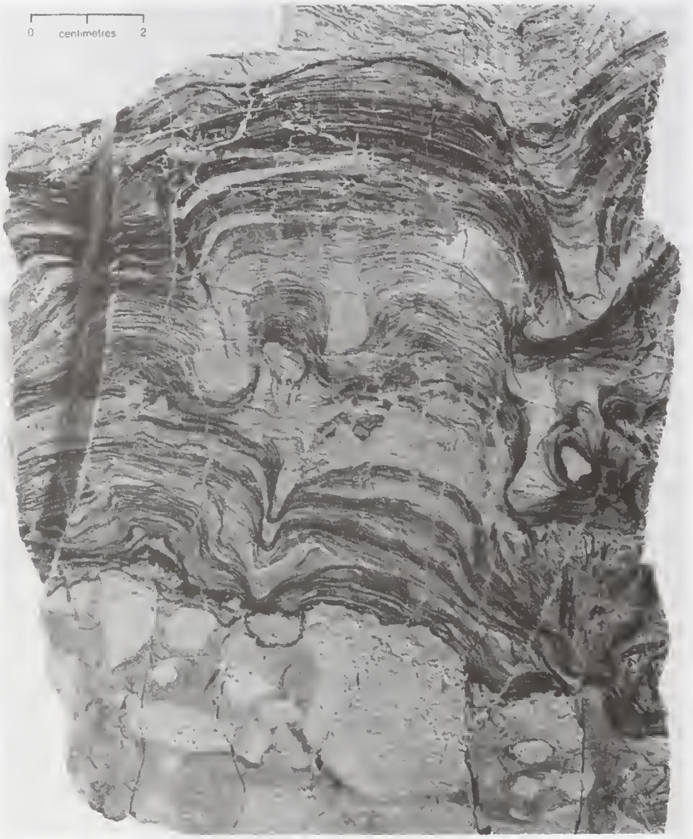


Figure 16: Stromatolites overlying the brecciated top of an ultramafic lava flow in Barberton Mountain Land, South Africa. The domal forms along the base are covered by larger, somewhat asymmetrical domes (see text).

By courtesy of Gary R Byerly, Louisiana State University, Baton Rouge

capable of cell division, which allowed DNA (deoxyribonucleic acid), the genetic coding material, to be passed on to succeeding generations.

By Early Proterozoic time both microfossils and stromatolites had proliferated. The best-known occurrence of microorganisms is in the two-billion-year-old, stromatolite-bearing Gunflint iron formation in the Huronian Basin of southern Ontario. These microbial fossils include some 30 different types with spheroidal, filamentous, and sporelike forms up to about 20 micrometres across. Sixteen species in 14 genera have been classified so far. Microfossils of this kind are abundant, contain beautifully preserved organic matter, and are extremely similar to such present-day microorganisms as blue-green algae and microbacteria. There are comparable microfossils of the Early Proterozoic in Minnesota and Michigan in the United States, the Belcher Islands in Hudson Bay in Canada, southern Greenland, Western Australia, and northern China. These microbiota lived at the time of the transition from an anoxygenic to an oxygenic atmosphere.

During the Late Proterozoic stromatolites reached their peak of development and became distributed worldwide. The first metazoa (multicelled organisms whose cells are differentiated into tissues and organs) also appeared at this time. The stromatolites diversified into complex, branching forms. From about 700 million years ago, however, they began to decline significantly in number. Possibly the newly arrived metazoa ate the stromatolitic algae, and their profuse growth destroyed the habitats of the latter.

Ediacaran fossils

Metazoa developed rapidly from the beginning of the Cambrian, when they acquired protective shells and hard skeletons, which contributed to their preservation in fossil form. More primitive metazoa without skeletons, however, appeared before that time—at the outset of the Ediacaran period about 700 to 670 million years ago. The type locality for these remarkable organisms is the Ediacara Hills in the Flinders Range north of Adelaide in South Australia, where there are an enormous number of well-preserved impressions in shallow-water quartzite stratigraphically situated some 500 metres below the base of the Cambrian System (see below *Cambrian life*). These are impressions of soft-bodied organisms that resemble modern jellyfish, worms, sponges, and sea pens, among which more than 60 species have been named. Comparable impressions are known from many parts of the world in the youngest Precambrian sediments, such as those at Charnwood in central England and in Ukraine, Siberia in Russia, Namibia, southeastern Newfoundland in Canada, and North Carolina in the eastern United States. Finally, there is the intriguing question as to when sexual division arose in life-forms. The American paleobiologist J. William Schopf has pointed out that in the abundant microflora of the 900-million-year-old Bitter Springs Formation of central Australia, some eukaryotic algae have cells in various stages of division into tetrahedral sporelike forms. These resemble the tetrad of spore cells of living plants known to develop by sexual division. In effect, by the end of the Precambrian the conditions were set for the explosion of life at the start of the Phanerozoic that ultimately led to the emergence of humankind. (B.F.W.)

Paleozoic Era

Major divisions

The Paleozoic (from the Greek for ancient life) is bounded by major events in the history of life. It began about 540 million years ago with an extraordinary diversification of marine animals and ended about 245 million years ago with the greatest extinction event in Earth history. The major divisions of the Paleozoic Era, from oldest to youngest, are the Cambrian, Ordovician, Silurian, Devonian, Carboniferous, and Permian periods. Some geologists recognize the Mississippian and Pennsylvanian periods in place of the Carboniferous (see Table 4).

Paleozoic rocks are widely distributed on all continents. Most are of sedimentary origin, and many show evidence of deposition in or near shallow oceans. Among the more useful guide fossils for correlation are trilobites (Cambrian to Ordovician), graptolites (Ordovician to Silurian), conodonts (Ordovician to Permian), ammonoids (Devonian to Permian), and fusulinids (Carboniferous to Permian).

On a global scale, the Paleozoic was a time of continental assembly. Cambrian continents were scattered, but none covered either pole, and the average world climate was probably warmer than it is today. The continent of Laurentia, composed mostly of present-day North America and Greenland, lay across the equator and remained there even after joining with other continents. By Ordovician time the large continent of Gondwana, consisting primarily of present-day Africa, Antarctica, Australia, South America, southern Europe, much of the Middle East, and India, began to move over the South Pole. The distribution of extensive glacial deposits has been used to track the movement of parts of Gondwana over and around the South Pole during the remainder of the Paleozoic. Parts of Gondwana, because of its large size, also extended into tropical latitudes. Siberia, essentially the large Asian part of what is now Russia, was a separate continent during the early and middle Paleozoic, when it moved from equatorial to north temperate latitudes. Baltica, composed mostly of present-day northern Europe (including Scandinavia), moved across the equator from southern cool temperate into northern warm latitudes during the Paleozoic, and it collided with and joined Laurentia during early Devonian time. Continued tectonic plate movements resulted in the final assembly of the supercontinent of Pangaea by the end of the Paleozoic. Such mountainous regions as the Appalachians, Caledonides, and Urals were originally deformed by the Paleozoic collision of the lithospheric plates. Large areas of all continents were episodically inundated by shallow seas, with the greatest inundations during the Ordovician and early Carboniferous periods.

At the beginning of the Paleozoic, animals were restricted to the oceans, and land plants had not appeared. About half of all animal phyla, especially those with hard shells and mineralized skeletons, originated during the early and middle Cambrian. The biota rapidly diversified throughout the Cambrian and Ordovician as life-forms adapted to virtually all marine environments. In numbers of described marine species, trilobites are the dominant kind of fossil in Cambrian rocks, whereas brachiopods predominate in Ordovician to Permian rocks.

Conquest of the land by life-forms

Several different kinds of organisms independently adapted to living on land, primarily during the middle Paleozoic. Leafless vascular plants (psilophytes) and invertebrate animals (centipede-like arthropods) were both established on land at least by Silurian time. Vertebrate animals made the transition to land via evolution of amphibians from air-breathing crossopterygian fish during the Devonian. Further conquest of the land became possible during the Carboniferous as dependence on moist environments for depositing spores and shell-less eggs was overcome, as plants evolved seeds (seed-fern origin), and as animals evolved amniote eggs with protective shells (reptile origin). Flight was first achieved also during the Carboniferous, as insects evolved wings.

The great extinction event at the end of the Paleozoic Era eliminated such major invertebrate groups as the blastoids, fusulinids, and trilobites. Other major groups, as, for example, the ammonoids, brachiopods, bryozoans, corals, and crinoids, were severely decimated but managed to survive. It has been estimated that as many as 95 percent of the marine invertebrate species perished during the late Permian. Extinction rates were much lower among vertebrates, both aquatic and terrestrial, and among plants. Causes of the extinction are not clear, but they may be related to changing climate and exceptionally low sea level. Although of lesser magnitude, other important mass extinctions occurred at the end of the Ordovician and during the late Devonian.

CAMBRIAN PERIOD

General considerations. The Cambrian, the earliest time division of the Paleozoic Era, extended from about 540 to 505 million years ago. Rocks formed or deposited during this time are assigned to the Cambrian System, which was named in 1835 by Adam Sedgwick for successions of slaty rocks in southern Wales and southwestern England. The corresponding period and system names are derived from Cambria, the Roman name for Wales.

As originally described, the Cambrian System was overlain by the Silurian System, which was named, also in 1835, by Roderick I. Murchison. Subsequent disagreement between Sedgwick and Murchison over the definition and placement of the Cambrian–Silurian boundary led to a bitter controversy that involved many British geologists. The problem persisted until after the deaths of both Sedgwick and Murchison in the 1870s and the eventual adoption of an intervening system, the Ordovician, which was proposed in 1879 by Charles Lapworth.

Rocks in the Cambrian-type area were divided by Sedgwick into what he called Lower, Middle, and Upper Cambrian. These rocks, however, are so poorly exposed, structurally complicated, and sparsely fossiliferous that they have had little influence on development of modern concepts of the Cambrian and its subdivisions. In fact, much of the type Cambrian has been reassigned to either the Precambrian or the Ordovician. Rocks in Wales that are now assigned to the Cambrian System roughly correspond to Sedgwick's Lower Cambrian.

Boundaries and subdivisions. The lower boundary of the Cambrian System is defined at a formal global stratotype section and point (GSSP), which was ratified by the International Union of Geological Sciences (IUGS) in 1992. The stratotype section is located at Fortune Head on the Burin Peninsula of southeastern Newfoundland in Canada. It contains a thick and continuous marine succession of mostly shale, siltstone, and sandstone. The stratotype point, representing a moment in time, is in the lower part of the Chapel Island Formation. It coincides with the base of the *Phycodes pedum* biozone and is close to the lowest stratigraphic occurrence of diverse shelly fossils. Many supplementary sections were investigated at boundary localities around the world before the stratotype was selected in Newfoundland. All these supplementary sections are important references for reconstructing the physical and biological histories of the boundary interval. Among the more thoroughly studied supplementary sections are those along the Aldan River of eastern Siberia and near K'un-ming in the Yunnan province of southern China.

The lower boundary of the Ordovician System indirectly defines the upper boundary of the Cambrian System. A formal boundary stratotype has not been selected, but work toward its selection is at an advanced stage. A working group sponsored by the IUGS has agreed to select a boundary stratotype close in age to the base of the Tremadoc Series, which has its type area in northern Wales. British geologists have traditionally assigned rocks and fossils of Tremadoc age to the Cambrian, whereas many others have assigned them to the Ordovician (see below *Ordovician Period: General considerations*). The stratotype for the Cambrian–Ordovician boundary is expected to be placed at the base of one of three closely spaced conodont zones and near the first appearance of planktonic (floating) graptolites. In rocks of this age, conodonts (toothlike microfossils produced by an extinct group of small, free-swimming marine animals) are among the best guide fossils for global time correlation, and planktonic graptolites have been used in defining the base of the Tremadoc Series and for zonation of the Ordovician System.

Cambrian series and stages

In most regions of the world, Cambrian rocks have been divided into lower, middle, and upper series (Table 5). The series boundaries, however, are not necessarily synchronous, because of differences in definition as well as problems in correlation. Some series have been further divided into stages, but these are mostly identifiable only within individual regions.

Cambrian rocks have a special biological significance, because they are the earliest to contain diverse fossils of animals. These rocks also include the first appearances of most animal phyla that have fossil records. Proliferation of organic lineages is called adaptive radiation, but Cambrian evolution produced such an extraordinary array of new body plans that the event has been referred to as the "Cambrian explosion." The beginning of this remarkable adaptive radiation has been used to divide the history of life on Earth into two unequal eons. The older, approximately 3-billion-year Cryptozoic Eon began with the appearance of life on Earth, and it is represented by rocks with mainly bacteria, algae, and similar primitive organisms. The younger, approximately half-billion-year Phanerozoic Eon, which began with the Cambrian explosion and continues to the present day, is characterized by rocks with conspicuous animal fossils.

Economic significance. Cambrian rocks are of moderate economic importance, as they provide a variety of resources. For example, ore bodies rich in such metals as lead, zinc, silver, gold, and tungsten have secondarily replaced Cambrian carbonate rocks, especially in parts of North America and Australia. Other carbonate rocks have been widely used as building stone and for making lime and portland cement. Large Cambrian phosphorite deposits are major sources of agricultural fertilizer in northern Australia, southwestern China, and southern Kazakhstan. Other Cambrian resources in China are mercury, uranium, and salt. Eastern Russia also has salt deposits of Cambrian age, as well as those of bauxite, the chief commercial source of aluminum. Some oil fields in southern Siberia produce oil from Lower Cambrian rocks.

Cambrian rocks. *Types and distribution.* Rocks of Cambrian age occur on all of the continents, and individual sections may range up to thousands of metres thick. The most fossiliferous and best-studied deposits are principally from marine continental-shelf environments. Among the thicker and better-documented sections are those in the Cordilleran region of western North America, the Siberian Platform of eastern Russia, and areas of central and southern China. Other well-documented, fossiliferous, but thinner sections are in Australia (especially western Queensland), the Appalachian Mountains of eastern North America, Kazakhstan, and the Baltic region (notably Sweden).

Lateral changes in the composition of Cambrian rocks resulted from regional differences in environments of deposition. Nearshore deposits are commonly composed of

Table 5: Subdivisions of the Cambrian System

system	series	Great Britain		Scandinavia		Russia and Kazakhstan	China	Australia	North America
			Tremadoc series					Datsonian	
Cambrian	Upper	Merioneth Series	no stages	*Olenus* Series	no stages	Aksayan	Fengshanian	Payntonian	Trempealeauan
						Saksian	Changshanian	unnamed stage	Franconian
						Ayusokkanian		Idamean	
							Kushanian	Mindyallan	Dresbachian
	Middle	Saint David's Series	no stages	*Paradoxides* Series	*Forchhammeri*	Mayan	Changhian	Boomerangian	no stages
					Paradoxissimus			Undillan	
						Amgan	Hsuchuangian	Floran	
					Oelandicus			Templetonian	
							Maochuangian		
	Lower	Comley Series	no stages	*Holmia* Series	no stages	Toyonian	Lungwangmiaoan	Ordian	no stages
						Botomian	Tsanglangpuian		
						Atdabanian	Chiungchussuian		
						Tommotian	Meishucunian		
						Manykaian			
						Vendian	Sinian	Ediacaran	

Development of carbonate platforms

siliceous sandstone. This usually grades seaward into siltstone and shale, which formed by accumulation of finer-grained sediment in deeper water where the seafloor was less affected by wave action. Extensive carbonate platforms, analogous to the modern Bahama Banks, developed along some continental shelves that were in low latitudes during Cambrian time. Rapid production of carbonate sediment in this warm, shallow-water environment resulted in massive deposits of Cambrian limestone and dolomite. Examples are exposed in the Cordilleran region of North America, in north-central Australia, along the Yangtze River in central China, and along the Lena River emanating in Siberia. Few Cambrian rocks from land environments have been documented, and most of those are of limited areal extent. They mainly represent deposits of floodplains and windblown sand. Without plants or animals, the desolation of Cambrian landscapes must have rivaled that of any present-day desert. In the absence of plants with roots to hold soil in place, Cambrian lands in general probably eroded more rapidly than they do now.

Probable causes of the rise in Cambrian sea level

Relative sea level rose significantly during the Cambrian, but with fluctuations. This is indicated by both the geographic distribution and the stratigraphic succession of sedimentary deposits. In North America, for example, early Cambrian marine deposits covered only marginal areas, but late Cambrian marine deposits covered much of the continent. A similar distribution of marine rocks is present on other continents. In stratigraphic sections from continental shelves that were in low latitudes it is common for a basal, nearshore sandstone to be overlain by a transgressive succession of more seaward shale and carbonate rocks. Shelf sections from high latitudes may be mostly or entirely sandstone, or a basal sandstone may grade upward into shale, but most of these sections contain evidence of marine transgression. Exceptions to the general Cambrian sea-level pattern are commonly attributable either to local tectonism or to different rates of sediment accumulation. The most likely explanation fór the general rise in Cambrian sea level seems to be increased thermal activity and related swelling of spreading ridges between lithospheric plates, which would displace vast quantities of seawater. It has been suggested that the general Cambrian transgression exerted an influence on adaptive radiation by greatly increasing the area of shallow seas where life was most abundant.

Correlation. Time correlation of Cambrian rocks has been based almost entirely on fossils. The most common fossils in Cambrian rocks are trilobites (Figure 17), which evolved rapidly and are the principal guide fossils for biostratigraphic zonation in all but rocks below the Atdabanian Stage or those of equivalent age (Table 5). Until the mid-1900s almost all trilobite zones were based on members of the order Polymerida. Such trilobites usually have more than five segments in the thorax, and the order includes about 95 percent of all trilobite species. Most polymeroids, however, lived on the seafloor, and genera and species were mostly endemic to the shelves of individual Cambrian continents. Therefore, polymeroid trilobites are useful for regional correlation but have limited value for intercontinental correlation, which has been difficult and subject to significant differences in interpretation.

Use of agnostoid fossils for intercontinental correlation

From the 1960s, investigators began to recognize that many species of the trilobite order Agnostida have intercontinental distributions in open-marine strata. These trilobites are small, rarely exceeding a few millimetres in length, and they have only two thoracic segments. Specialized appendages, which were probably useful for swimming but unsuitable for walking on the seafloor, suggest that they were pelagic. Agnostoids make up less than 5 percent of all trilobite species, but individuals of some agnostoid species are abundant. This fact, together with their wide geographic distribution and rapid evolution, makes them valuable for refined intercontinental correlation. Agnostoids first appear in upper Lower Cambrian rocks but did not become common or diversify significantly until the middle of the Cambrian. Therefore, agnostoids have their greatest biostratigraphic value in the upper half of the Cambrian System. A comprehensive trilobite zona-

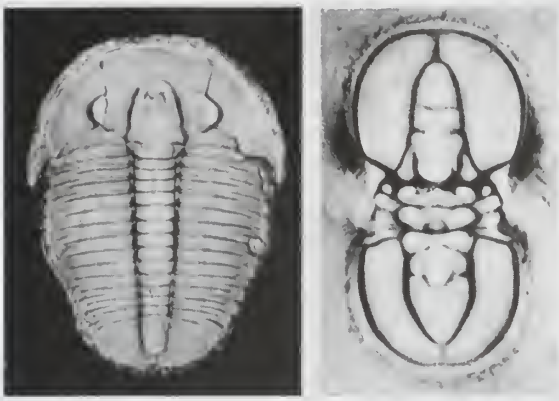


Figure 17: *Representative Cambrian trilobites.* (Left) *Elrathia kingii* (order Polymerida) and (right) *Ptychagnostus gibbus* (order Agnostida).

By courtesy of R A Robison

tion in Sweden has frequently been cited as a standard for correlation.

Other kinds of fossils have had more limited use in Cambrian biostratigraphy and correlation. Among them are the archaeocyathan sponges in the Lower Cambrian and brachiopods throughout the Cambrian, but use of both groups has been hampered by problems of endemism. Small mollusks and other small shelly fossils, mostly of problematic affinities, have been employed for biostratigraphy in the Tommotian Stage, but their utility is also limited by endemism. Conodonts appear in the uppermost Precambrian but are rare in most Cambrian rocks except those of latest Cambrian age, when adaptive radiation of conodont animals accelerated. Wide species distributions, rapid evolution, and abundance make conodonts excellent indexes for global biostratigraphy in uppermost Cambrian to uppermost Triassic rocks.

Since roughly the 1980s, trace fossils have been used with limited precision to correlate uppermost Precambrian and basal Cambrian strata. Although the biostratigraphic use of such fossils has many problems, they nevertheless demonstrate progressively more complex and diverse patterns of locomotion and feeding by benthic (bottom-dwelling) marine animals. For example, *Phycodes pedum,* which initially appears in basal Cambrian deposits, is the first regularly branching burrow pattern.

Cambrian environment. *Paleogeography.* The geography of the Cambrian world differed greatly from that of the present. The geographic reconstruction in Figure 18 is based on integrated geologic and biological evidence. Fossils in continental-shelf deposits indicate the presence of at least three major faunal provinces during much of the Cambrian Period.

The most distinct faunal province surrounded the continent of Laurentia. Paleomagnetic evidence indicates that Laurentia was located over the equator during most or all of Cambrian time. This geographic interpretation is supported by the presence of thick, warm-water, carbonate-platform deposits that accumulated in a broad belt encircling the continent. These carbonates are commonly flanked on the inner shelf by lagoonal shale and nearshore sandstone. On the outer shelf, the carbonates commonly grade into laminated mudstone and shale that accumulated in deeper water. At times, two almost mutually exclusive subfaunas were separated by temperature and salinity barriers in the shallow water on the carbonate platforms. Inner, restricted-shelf deposits are characterized by sparse, low-diversity faunas that tend to be highly endemic. Outer, open-shelf deposits are characterized by common to abundant, high-diversity faunas that are widely distributed around the continent. Fossils are usually most abundant and most diverse near the outer margin of the carbonate platform. Because Laurentia has remained nearly intact structurally, it is ideal for studying the relationships between Cambrian environments and faunas around a low-latitude Cambrian continent.

Another Cambrian faunal province surrounded the small continent of Baltica, which was located in middle to high southern latitudes. Cambrian shelf deposits of Baltica are relatively thin, rarely exceeding 250 metres in thickness, and are composed of primarily sandstone and shale. Seemingly as a consequence of cool-water environments, carbonate deposits are relatively minor and very thin. The wide distribution of many species from the nearshore to deep-shelf environments of Baltica suggests no significant restriction in shelf dispersal like that caused by the shallow carbonate platforms of Laurentia.

Largest Cambrian faunal province

The largest Cambrian faunal province is that around the continent of Gondwana, which extended from the low northern latitudes to the high southern latitudes, just short of the South Pole. Rocks and faunas of Gondwana show major changes that correspond to its great size and wide range of climates and environments. The Antarctic and Australian sectors of Gondwana were in low latitudes and have extensive carbonate deposits, although those of Antarctica are poorly exposed through the present-day polar ice cap. Faunal differences and paleomagnetic evidence suggest that present-day North and South China were on separate tectonic plates. Extensive carbonate deposits in both regions, however, indicate that both plates were in low latitudes. South China has strong faunal similarities to both Australia and Kazakhstan, but details of the Cambrian geographic relationships remain unclear.

Several terranes seem to have been near or attached along the margin of Gondwana in high southern latitudes (the northern Africa sector), but many details of their Cambrian geographic relations are unknown. (Terranes are fault-bounded fragments of the Earth's crust characterized by a geologic history markedly different from that of neighbouring crustal segments.) These terranes now make up much of southern Europe and parts of eastern North America. Cambrian deposits in all the terranes are chiefly sandstone and shale and include few or no carbonates. Their faunas closely resemble those of Baltica at generic and higher taxonomic levels, but differences at the species level suggest some geographic separation.

Siberia was a separate continent located in the low latitudes between Laurentia and Gondwana. Faunal affinities suggest that it lay relatively close to equatorial Gondwana.

Present-day Kazakhstan seems to be composed of several microcontinental blocks that were in all likelihood separated during the Cambrian. These blocks were amalgamated after the Cambrian, and the composite continent, Kazakhstania, collided with Siberia during the late Paleozoic.

Few Cambrian faunas from continental-slope and deep-ocean environments are known. Limited information from these is important, however, for demonstrating affinities between deep, cool-water faunas from all latitudes and shallow, cool-water faunas from high latitudes. Close similarity between the observed distribution patterns of Cambrian and modern marine arthropods has been used as persuasive evidence for thermally stratified Cambrian oceans in lower latitudes and for a thermocline separating warm-water and cool-water layers (Figure 19). The inferred thermocline as well as wide oceanic separation were likely causes for the high endemism of the Laurentian faunal province. This interpretation is supported by Middle Cambrian deposits in North Greenland where, in a few tens of kilometres, normal Laurentian shelf-margin trilobite faunas grade into deepwater faunas like those in the shallow-shelf deposits of Baltica. In a similar pattern, trilobite species in deepwater faunas of Late Cambrian age in the western United States and southeastern China are the same, but shallow-water faunas of the same age in the two regions have few genera in common.

As large lithospheric plates continued to move during the Phanerozoic Eon, terranes of various sizes were displaced. Endemic Cambrian fossils, in conjunction with such other geologic evidence as physical stratigraphy, have been useful in helping to identify the geographic origins of some terranes, particularly those that have undergone substantial displacement. Examples are northern Scotland with Laurentian faunas, eastern Newfoundland with Baltic faunas, southern Mexico (Oaxaca) with Gondwanan (South American) faunas, and west-central Argentina (Precordillera) with Laurentian faunas.

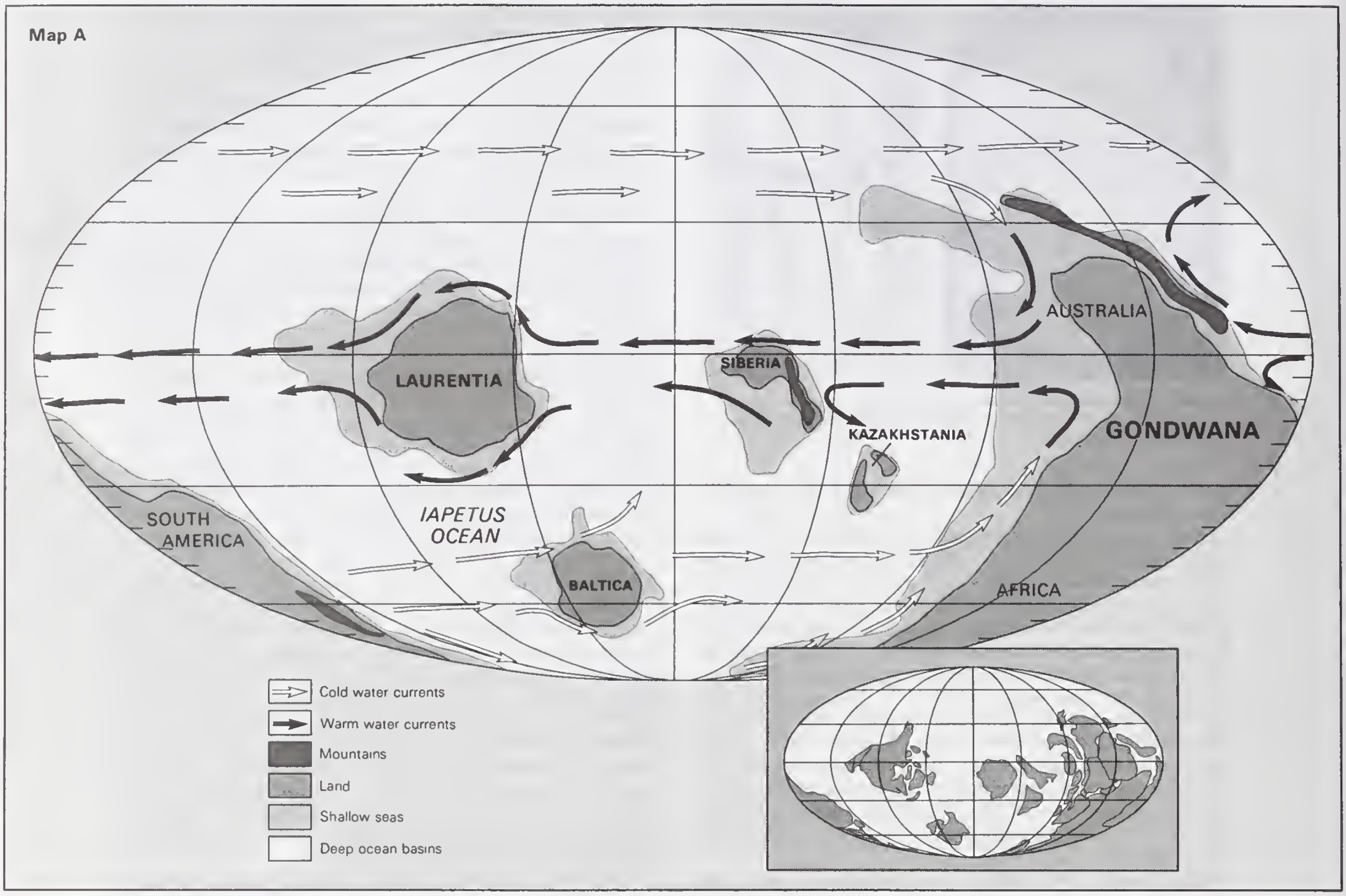


Figure 18: Distribution of landmasses, mountainous regions, shallow seas, and deep ocean basins during the Late Cambrian. Included in the paleogeographic reconstruction are cold and warm ocean currents. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset at the lower right. Map B provides a "backside" view of the reconstruction shown in Map A.

Adapted from C R Scotese, The University of Texas at Arlington

Another important consequence of continued plate movement has been the formation of large mountain ranges by crumpling where plates have collided. Pressure and heat generated during collisions since the Cambrian have folded, faulted, and metamorphosed significant volumes of Cambrian rock, especially that from the outer margins and slopes of many continental shelves. No crustal rocks in today's oceans have been found to be older than Mesozoic in age, and apparently most pre-Mesozoic deposits that accumulated in the deep ocean basins have been destroyed by subduction into the Earth's interior.

Abrupt global changes in sea level

Relatively abrupt changes in sea level may have significantly influenced Cambrian environments and life. A global drop in sea level is suggested by extensive unconformities (*i.e.,* interruptions in the continuity of depositional sequence) and changes in sedimentary rocks near the Early–Middle Cambrian boundary. The time represented by such unconformities in sectors of Laurentia and Baltica bounding the Iapetus Ocean has been called the Hawke Bay event. An apparent absence of a coeval unconformity in western North America seems to be an anomaly. Thick, uninterrupted shelf deposits in this sector of Laurentia, however, may have resulted from abnormal shelf subsidence caused by cooling of crustal rocks following a late Precambrian plate-rifting event. Temporal correlations with unconformities on other continents lack precision. Nevertheless, it is perhaps significant that a number of characteristic Early Cambrian animal groups were either exterminated or severely restricted in their geographic distribution at about the same time in the world's shallow-shelf environments. Among biostratigraphically important trilobites, the olenellids were exterminated around Laurentia, the holmiids were killed off around Baltica, and the redlichiids vanished around Gondwana. Also, diverse and abundant reef-dwelling archaeocyathans disappeared from most low-latitude, warm-water continental shelves.

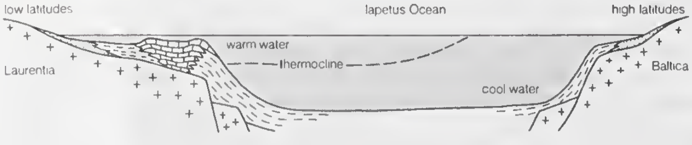


Figure 19: Hypothetical cross section of the Iapetus Ocean during Middle Cambrian time (vertical scale greatly exaggerated).

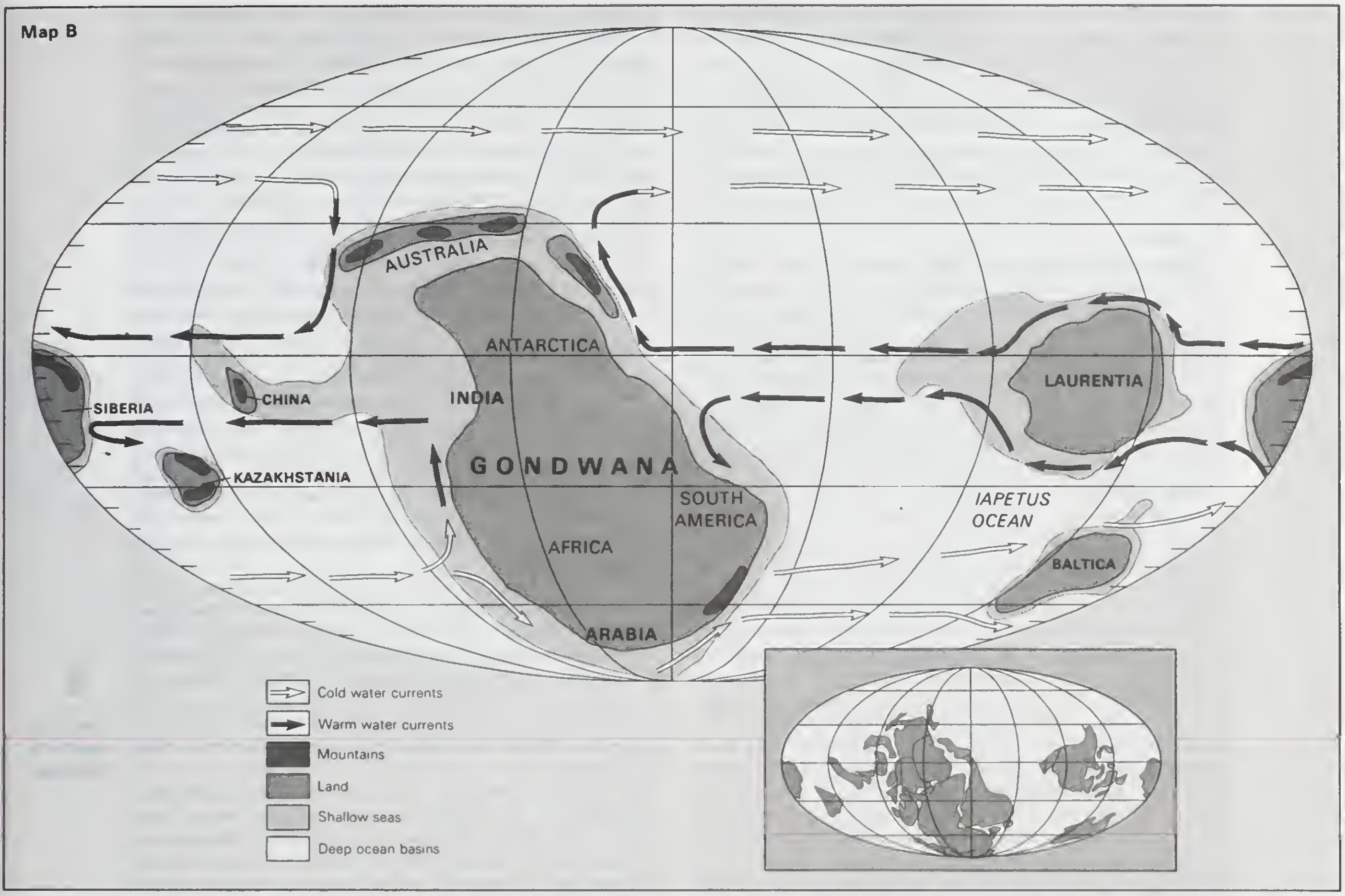
A significant rise in sea level is suggested by rather abrupt and extensive displacements in sedimentary environments and biotas in the middle Middle Cambrian (the *Ptychagnostus gibbus* zone). Lowland areas were flooded, as in parts of Baltica. In warm-water shelf sections of the world, it is common for coarse-grained, shallow-water, carbonate rocks to be abruptly overlain by fine-grained, deeper-water, laminated limestone or shale. Adaptive radiation of the pelagic agnostoid trilobites was greatly accelerated in open-oceanic environments following this event, perhaps in response to expanded habitats.

Missing faunas and an unconformity, which define the boundary between the Dresbachian and Franconian stages in peripheral areas of North America, suggest another significant drop in sea level during the early Late Cambrian. Evidence for two other lesser changes in Cambrian sea level has been identified near the Cambrian–Ordovician boundary. Associated minor unconformities have provided a problem for selection of a boundary stratotype, which ideally should be located in an uninterrupted stratigraphic section.

Sites of volcanic activity

Several regions of Cambrian volcanism have been identified. Australia was especially active, with large areas in the

Map B



northern and central regions covered by flood basalts during the Early Cambrian and with residual activity into the Middle Cambrian. Basalts and mafic intrusives in southeastern Australia formed in a volcanic island-arc setting during the Early and Middle Cambrian. Volcanic suites of similar age also are present in New Zealand and in parts of Antarctica (northern Victoria Land, Ellsworth Mountains, and Pensacola Mountains). Other significant Lower and Middle Cambrian volcanic deposits are present in southern Siberia and western Mongolia (Altai and Sayan mountains), eastern Kazakhstan and northwestern China (Tien Shan), and northeastern China. Cambrian volcanics are scattered along the easternmost margin of the United States, but most are probably island-arc deposits that were accreted to Laurentia after the Cambrian. In the southern United States (Oklahoma), granitic intrusives and basaltic and rhyolitic extrusives are associated with a large tectonic trough that was formed by Early and Middle Cambrian crustal extension.

Minor volcanic deposits, mainly ash beds and thin flows, are widely known. In general, these have received little study, but some are suitable for determination of isotopic ages. Zircons from a lower Lower Cambrian (pre-Tommotian) volcanic ash bed in New Brunswick, Can., have a uranium-lead age of 531 million years. Volcanic tuffs near inferred Tommotian–Atdabanian boundaries in both Morocco and southwestern China have yielded similar dates of 521 million years.

The tectonic history of the Paleozoic is much better known than that of the Precambrian. In general, however, late Precambrian history seems to have been characterized by continental fragmentation, whereas Paleozoic history was characterized by continental accretion. The Cambrian was a period of transition between those tectonic modes, and continents were scattered, apparently by fragmentation of a late Precambrian supercontinent. Major Cambrian and early Ordovician tectonism affected large areas of Gondwana in what are now Australia, Antarctica, and Argentina. Multiphase tectonism in Antarctica is called the Ross Orogeny, and in Australia it is known as the Delamerian Orogeny. At least some of the volcanic activity noted above, particularly that of volcanic island arcs, is evidence that seafloor spreading and crustal subduction were active geologic processes.

Paleoclimate. Global climate during Cambrian time was probably warmer and more equable than it is today. An absence of Cambrian glacial deposits and an abundance of widespread, warm-water, carbonate deposits both suggest higher average temperatures than at present. The absence of glacial deposits of Cambrian age is more notable because such deposits are common and widespread in the upper Precambrian, and they accumulated again during the Ordovician in northern Africa as Gondwana began to move over the South Pole. An apparent absence of either land or landlocked seas at the Cambrian poles may have prevented the accumulation of polar ice caps.

Cambrian life. The long history of life on Earth has been punctuated by relatively abrupt changes. Some have argued that the greatest change of all occurred in marine environments near the Precambrian–Cambrian boundary. Fossils from Cambrian rocks include the oldest representatives of most animal phyla having mineralized shells or skeletons. A lack of observed connecting links suggests that processes of biomineralization evolved independently in several phyla. Whether or not soft-bodied representatives of some of these phyla originated during the Precambrian but have no preserved record is a debated question. Nevertheless, the hard parts of Cambrian animals had a much greater potential for preservation than soft parts, and they mark the beginning of a diverse fossil record.

Fossil record of the Precambrian–Cambrian transition

Preservation of the record of the Precambrian–Cambrian transition was significantly affected by global changes in sea level. During latest Precambrian time, the sea level was relatively low, resulting in areally restricted oceans and expanded continents. Throughout much of the Cambrian, rising seas gradually flooded vast land areas. Sediment was eroded from the continents and deposited in adjacent seas. Because of low sea level, the sedimentary and fossil records of the Precambrian–Cambrian transition are generally most complete toward the outer margins of continental shelves. As a corollary, the time gap represented by the boundary surface generally increases in landward directions. Absence or serious incompleteness of a transitional record in most areas, particularly in those of classical Cambrian studies, contributed significantly to the long-held notion of an abrupt or sudden appearance of Cambrian fossils. This was compounded by a general deficiency in knowledge of Precambrian biotas before the mid-1900s.

Considering the biological importance of the Precambrian–Cambrian transition, it is somewhat surprising that the primary impetus for its detailed study came from a project undertaken to establish international agreement on a suitable boundary stratotype. Before the project was initiated in 1972, reasonably complete stratigraphic sections across the transition were either largely unrecognized or ignored. Since 1972, information about the transition has accumulated at an accelerating rate. Although many details remain to be learned, the general history of this momentous interval is becoming clear.

The Precambrian–Cambrian biotic transition, once thought to be sudden or abrupt, has been found to include a succession of events spread over many millions of years. It commences with the appearance of the animal kingdom (*i.e.,* multicelled organisms that ingest food), but the date and details of that event remain obscure. At least three informal phases in the transition can be identified by progressively more diverse and complex biotas.

The earliest phase, of late Precambrian age, is characterized by fossils of soft-bodied animals known from many localities around the world. These organisms may have appeared as early as 650 million years ago and are commonly called the Ediacaran fauna (see above *Precambrian time: Precambrian life*). The fossils are predominantly the imprints of soft-bodied animals. Their extraordinary preservation, usually in sandstone or shale, was probably the result of rapid burial and protection by smothering sediment. Most of the fossils are relatively simple, and many resemble worms, sea pens, and jellyfish. Dwelling traces like those of modern sea anemones are also common. Higher taxonomic assignments are controversial, however, because critical diagnostic features are not evident. Some paleontologists have assigned Ediacaran body fossils to the extant phyla Annelida, Coelenterata, and Arthropoda, whereas others have regarded them as members of extinct taxonomic groups of high rank. Some adherents of the latter viewpoint have suggested that the Ediacaran fauna was terminated by a major extinction event, but direct evidence of an abrupt faunal replacement has not been detected in any stratigraphic section.

Other kinds of fossils also provide valuable clues about life during Ediacaran time. Photosynthetic organisms include unicellular blue-green algae (cyanobacteria) and acritarchs (probable algae), both of low diversity. Individuals of some species were probably abundant, however, and may have been an important source of food for Ediacaran animals. Hard parts of animals, primarily known from Africa and China, are mainly dwelling tubes composed of calcium carbonate and other compounds. Most were probably secreted by sessile, filter-feeding, wormlike animals. Although rare and of low diversity, these forms are significant because they signal the advent of biomineralization. The oldest unequivocal trace fossils, mainly crawling trails, are also of Ediacaran age. The trails suggest that locomotion of the trace makers was accomplished by waves of muscular contraction, like that in annelids and sea slugs, and not by legs. All but the latest Ediacaran trace fossils are relatively simple, suggesting limited and primitive behaviour patterns. Their low diversity further suggests that few kinds of mobile animals lived on the Ediacaran seafloor.

Oldest trace fossils

The second phase of the Precambrian–Cambrian biotic transition is characterized by a marked increase in the diversity of its shelly fauna and a lack of trilobites. It is near the lowest stratigraphic occurrence of this fauna that the Precambrian–Cambrian boundary stratotype has been placed. The fauna includes that of the Tommotian Stage, as applied in Russia, and it has often been referred to as the Tommotian fauna. It is known from many localities around the world, but time correlations lack precision. A general acceleration in biotic diversity during this second phase is the beginning of the so-called Cambrian explosion.

Fossils of the second phase, which may be locally abundant, represent several new animal groups of Paleozoic aspect. Calcified archaeocyathan sponges diversified rapidly and were the first skeletal metazoans to develop a modular growth habit. They also evolved a complex symbiotic relationship with reef-building blue-green algae. Mollusks, preserved in both shale and limestone, include at least four classes (Monoplacophora, Gastropoda, Hyolitha, and Rostroconchia). Brachiopods made their appearance but are low in diversity. Several problematic groups are represented by an astonishing array of small mineralized tubes, scales, and spicules. The presence of arthropods, the first animals to develop legs, is indicated by characteristic trace fossils. The skeletal remains of arthropods are not preserved in the fauna, ˛however, presumably because they were not mineralized. Other trace fossils show a marked increase in abundance and diversity as well as an expansion of behaviour patterns that reflect improvements in locomotion, greater ability to penetrate sediment, and new foraging strategies.

The third phase of the Precambrian–Cambrian biotic transition commenced with the appearance of mineralized trilobite skeletons, which approximately correlates with the base of the Atdabanian Stage of the Lower Cambrian, as conceived in Russia. Subsequent adaptive radiation of the trilobites was exceptional, and their remains dominate most later Cambrian deposits. For this reason, the Cambrian Period has sometimes been called the "Age of Trilobites."

Predominance of trilobites

The known Cambrian biota was restricted to marine environments. At least 11 extant animal phyla (Annelida, Arthropoda, Brachiopoda, Chordata, Ctenophora, Echinodermata, Hemichordata, Mollusca, Onychophora, Porifera, and Priapulida), including most of those with a fossil record, first appear in Cambrian rocks. Most of these rapidly diversified as they seemingly adapted to numerous unfilled ecological niches. Another five phyla (Nemertea, Phoronida, Platyhelminthes, Pogonophora, and Sipuncula) are questionably known from Cambrian fossils. The only extant animal phylum with a good fossil record that is not known from Cambrian rocks is the Bryozoa, which first appears in rocks of Early Ordovician age. A summary of the principal biotic groups of the Cambrian is given below.

Cambrian photosynthetic organisms, the primary food of animals, are entirely unicellular. These organisms include a variety of bacteria and algae of the kingdoms Monera and Protista. Their evolution, like that in associated animals, shows a marked acceleration in adaptive radiation and biomineralization near the base of the Cambrian. A new calcareous benthic (bottom-dwelling) flora dominated by blue-green algae appeared. Some of these organisms formed mounds on the seafloor. Others formed small, concentrically laminated, marble- or biscuit-shaped structures called oncoids, which were locally abundant. Although it was rarely preserved, there existed a noncalcareous benthic flora that also was dominated by blue-green algae. By at least Middle Cambrian time, some noncalcareous green algae (Chlorophyta) had become common. In North America and Siberia, the axes of one species, *Margaretia dorus,* exceeded two centimetres in diameter and were probably more than one metre in height. Such large size is attained by modern green algae only in warm, equatorial oceans. The phytoplankton, consisting of acritarchs and blue-green algae, also diversified near the base of the Cambrian. Acritarchs are widespread in many kinds of marine rocks and seem to have potential for an improved zonation of Lower Cambrian rocks. They are difficult to study, however, because of their microscopic size.

Cambrian faunas, like those of the present day, are commonly dominated in numbers and kind by members of the phylum Arthropoda. Calcification of skeletons by the beginning of Atdabanian time contributed to an abundant fossil record of the class Trilobita, of which some details have been discussed above. Many hundreds of genera and thousands of species of Cambrian trilobites have been described worldwide. Rates of evolution in Cambrian trilobites were relatively rapid, resulting in short stratigraphic ranges and giving them much value for biostratigraphic correlation. Representatives of the class Ostracoda, characteristically enclosed by a bivalved carapace, also appeared near the base of the Atdabanian. Compared to trilobites, however, ostracods are generally rare and of low diversity throughout the Cambrian, except in some rocks of Australia and China. Extraordinary preservation at rare localities indicates that many other kinds of arthropods were at least locally more abundant and more diverse than the trilobites. These other arthropods had unmineralized skeletons, and some may represent extinct classes.

Abundance of sponges

Sponges (phylum Porifera) are commonly represented in Cambrian faunas. Archaeocyathan sponges, characterized by cup-shaped skeletons with double calcareous walls and numerous pores, are abundant and diverse in some Early Cambrian deposits. They have been used for provincial biostratigraphic zonation, especially in Australia and Siberia, and archaeocyathan taxa define all but the lowest one of the five Lower Cambrian stages in Siberia (Table 5). Archaeocyathans are common only in regions that were in low Cambrian latitudes, including Antarctica, Australia, China, Kazakhstan, Siberia, and North America. Their latitudinal distribution is similar to that of modern colonial corals, suggesting adaptation to similar ecological controls in warm shallow seas. Archaeocyathans nearly disappeared at the end of the Early Cambrian, but rare species survived until middle Late Cambrian time, after which the group became extinct. Other common Cambrian sponges had skeletons of siliceous spicules, which readily disaggregated after death, making their identification at lower taxonomic levels difficult, if not impossible. At rare localities where preservation is exceptional, including articulated skeletons and associated soft-bodied taxa, spicular sponges are second only to arthropods in species diversity. This suggests that Cambrian sponges were much more common and more diverse than is indicated by the known fossil record. Limited information indicates that species of spicular sponges evolved slowly during the Cambrian, resulting in relatively long stratigraphic ranges.

Brachiopod shells are present in many Cambrian continental-shelf deposits. In total number of species that have been described from Cambrian rocks, brachiopods are second only to trilobites. Species diversity, however, is generally low to moderate at most localities. Phosphatic shells of the class Inarticulata are normally much more common and more diverse than are calcareous shells of the class Articulata. These abundance and diversity relationships are usually reversed in post-Cambrian rocks.

The phylum Echinodermata (some present-day representatives of which are sea urchins and starfish) had a major adaptive radiation during the Cambrian Period. The number of classes increased from three in the Early Cambrian to eight in the Middle Cambrian. Only one of these, the Eocrinoidea, is known from many species, but the described record seems to be grossly incomplete. Skeletal plates in early echinoderms were not rigidly connected, and they readily disaggregated after the death of an animal. Consequently, it is rare to find articulated skeletons that can be identified at lower taxonomic levels. In some Cambrian limestones, however, skeletal plates of echinoderms are a dominant sedimentary constituent, indicating the existence of innumerable animals and suggesting far greater diversity, especially at low taxonomic levels, than has been recorded. As in some modern echinoderm species, it is common for those in the Cambrian to show evidence of a gregarious habit and patchy distribution. Most of the Cambrian echinoderms were suspension and detritus feeders, and it was only after the Cambrian that herbivores and carnivores became common. All classes of echinoderms that were present during the Cambrian, except the Crinoidea, subsequently became extinct.

Proliferation of mollusks

The phylum Mollusca also underwent significant adaptive radiation during the Cambrian, with the appearance of the classes Monoplacophora, Gastropoda, Pelecypoda (synonymous with Bivalvia), Cephalopoda, Polyplacophora, Rostroconchia, Hyolitha, and Stenothecoida. (The latter three are now extinct.) The only molluscan class that appeared after the Cambrian is the Scaphopoda (tusk or tooth shells), which originated during the Ordovician. A small variety of mollusks is present in the shelly fauna of the earliest Cambrian. Mollusk shells usually are absent or rare in later Cambrian rocks, but at a few localities they are common to abundant. The small conical shells of hyoliths are the kind most commonly preserved in Cambrian rocks.

Other new Cambrian phyla largely lack biomineralization and have a poor fossil record. The Hemichordata is represented by rare sessile graptolites (order Dendroidea) of the class Graptolithina, which appeared during the Middle Cambrian. Appearance of the more common planktonic graptolites (order Graptoloidea) has been informally used to define the Cambrian–Ordovician boundary, and the formal boundary stratotype is expected to be established close to this biohorizon. Cambrian worm phyla include the Annelida, Priapulida, and probable Pogonophora, but these are mainly known from localities where preservation was extraordinary. Other rarely represented phyla are the Onychophora, with leglike lobopodia, and Ctenophora (comb jellies).

The origin of the phylum Chordata is unclear. If primitive conodont-like fossils (paraconodonts) are included, as argued by some paleontologists, the phylum appeared during the late Precambrian. Rare, soft-bodied, possible chordates have been described from Lower Cambrian rocks. The oldest unequivocal chordate remains are isolated bony plates of jawless fish in Upper Cambrian rocks of the western United States.

Trace fossils, as discussed above, provide independent evidence of accelerated animal diversification and a distinct increase in the complexity of animal behaviour near the beginning of the Cambrian Period. Other evidence from trace fossils indicates changes in Cambrian bioturbation, the churning and stirring of seafloor sediment by animal forms. Late Precambrian (Ediacaran) trace fossils from around the world are in the main surface trails and show little evidence of sediment burrowing. Quantitative study in the western United States has shown that a significant increase in bioturbation occurs between pre-trilobite (Tommotian) and trilobite-bearing (Atdabanian) Lower Cambrian rocks. Throughout the Cambrian, bioturbation was more intensive in nearshore and inner-shelf environments than in more offshore settings. The depth of bioturbation in carbonate environments of the inner shelf was consistently less than a few centimetres throughout Cambrian time.

Cambrian deposits with soft-bodied organisms

Modern biotas are largely dominated by soft-bodied organisms, whereas the fossil record is overwhelmingly dominated by the hard parts of organisms. Rare deposits of fossils with soft parts are therefore of great importance in helping to establish the original diversity and ecology of ancient biotas. Among the most famous soft-bodied biotas is that in the Burgess Shale of western Canada (British Columbia), which is early Middle Cambrian in age. Tens of thousands of complete specimens, many with soft parts preserved in remarkable detail, were apparently buried by submarine slumping of sediment on the continental shelf of Laurentia. Fossils from the Burgess Shale have been used to demonstrate the presence of a complex community as diverse in habit, structure, and adaptation as many modern communities. If isolated, fossils with hard parts would constitute a typical Cambrian fauna, but they represent only about 40 percent of the genera in the Burgess Shale, a proportion similar to that in modern faunas on continental shelves.

Other less diverse Cambrian deposits with soft-bodied organisms have been discovered in such places as South Australia, China (Yunnan), North Greenland, Sweden, and the United States (Utah and Pennsylvania). Some of these are important in demonstrating that the biota of the Burgess Shale is unusual only in preservation and not in composition. They also demonstrate that some of the soft-bodied taxa have substantial geologic ranges and wide geographic distributions. Extraordinary preservation of Late Cambrian arthropods in Sweden is especially notable, as the bodies and appendages remain largely uncrushed and the integument retains many fine structures, including setae and pores.

Minor extinction events occurred sporadically throughout the Cambrian Period. One near the end of the Early Cambrian was apparently related to global marine regression. At least three Late Cambrian events primarily affected low-latitude shelf faunas and have been used in North America to define biostratigraphic units called biomeres. (Such units are bounded by sudden nonevolutionary changes in the dominant elements of a phylum.) Each of the Cambrian biomere events eliminated several trilobite families, which collectively contained most of the genera and species that were living on the continental shelves. Less attention has been paid to extinction patterns among other invertebrates, but some evidence of corresponding extinctions among brachiopods and conodonts is available. Geochemical evidence suggests that the biomere extinctions were probably caused by abrupt drops in water temperature. Oxygen isotopes from the skeletons of bottom-dwelling trilobites associated with one biomere boundary in Texas indicate a drop in water temperature of about 5° C (9° F) at the boundary. A comparable decrease in temperature would kill the larvae of many modern marine invertebrates that live in warm oceans. Following each Cambrian extinction, shelf environments were repopulated by low-diversity trilobite faunas of relatively simple form, which apparently emigrated from deeper and cooler off-shelf environments. In effect, every one of the biomere events was followed by an adaptive radiation of new taxa, especially among the trilobites. (R.A.R.)

ORDOVICIAN PERIOD

General considerations. The Ordovician, the second oldest period of the Paleozoic Era, is thought to have covered the span of time between 505 and 438 million years ago (see Table 4), though radiometric age determinations may range from as much as 515 to 435 million years ago. The rocks that originated during the period make up the Ordovician System. The designation Ordovician was proposed in 1879 by the English geologist Charles Lapworth, who derived it from Ordovices, the name of a Celtic tribe that had inhabited a part of North Wales at the time of the Roman invasion of Britain. Lapworth applied the term to a segment of rocks exposed in the Arenig mountains located roughly 40 kilometres west of the Welsh–English border and eastward to the Bala area of North Wales. This area constitutes a type area (or section), because the fossil faunas in the rocks are, as Lapworth recognized, characteristic of a unique geologic time interval. The stratigraphic position of the Ordovician rocks above those bearing fossil faunas typical of the Cambrian and beneath those bearing faunas representative of the Silurian is readily demonstrable. This section lies between an anticlinorium on the west in which Cambrian rocks are exposed and a synclinorium on the east in which fossiliferous Silurian rocks occur. (An anticlinorium is a large anticline on which minor folds are superimposed, while a synclinorium is a large syncline on which such folds are superimposed.)

Ordovician type area

Lapworth had in mind faunal aggregates as the basis for establishing not only the Ordovician but also the Cambrian and Silurian, stating that he saw

> three distinct faunas, as broadly marked in their characteristic features as any of those typical of the accepted systems of a later age. The necessity for tripartite grouping of Lower Palaeozoic rocks and fossils, in partial accordance with this fact, has been very generally acknowledged for the last thirty years.

With this discussion, Lapworth resolved a long-standing debate among British geologists concerning the division of the Lower Paleozoic rocks—those beneath the Devonian Old Red Sandstone. (R.I. Murchison had asserted that only the Silurian encompassed the Lower Paleozoic, while A. Sedgwick had proposed that the latter be divided into two periods, the Cambrian and the Silurian. For a fuller discussion of this controversy, see the sections *Cambrian Period* and *Silurian Period.*)

Trilobites, brachiopods, and graptolites were the major organisms considered in identifying Ordovician rocks. The distinctive brachiopod–trilobite–graptolite faunas regarded as characteristic of the Ordovician by Lapworth have been found in fossil-bearing rocks on all the continents. Notable, too, is the fact that Ordovician rocks have the distinction of being the topographically highest rocks on Earth. Fossiliferous Ordovician carbonates underlie the uppermost part of Mount Everest, the world's tallest mountain, located on the Nepal–Tibet border.

Ordovician series

Boundaries and subdivisions. In Great Britain the Ordovician System is divided into six series. From oldest to youngest, they are the Tremadoc, Arenig, Llanvirn, Llandeilo, Caradoc, and Ashgill. Each series is based on a group of rocks bearing characteristic fossils exposed in a type area. Type Tremadoc rocks are found around Tremadoc Bay in North Wales. Type Arenig rocks and faunas occur near Arenig Fawr in North Wales. The type Llanvirn occurs in sea cliffs and related exposures near Saint David's in South Wales. Llandeilo rocks and fossil faunas are evident in the environs of Llandeilo, Wales. Caradoc rocks and fossils occur near the Onny River in Shropshire, east of Wales. Ashgill strata and faunas occur in the English Lake District. Brachiopods and trilobites constitute the characteristic faunal aggregates of each series, except the Llanvirn in which graptolites are common.

These Ordovician series have been used widely in Europe in discussing Earth and life history during the Ordovician. Although attempts have been made to apply these divisions around the world, the faunas that typify each series appear to be limited to Europe in their distribution. Accordingly, alternative divisions of the Ordovician have been proposed in other countries (see Table 6).

In Lapworth's original definition, the base of the Arenig was the base of the Ordovician. However, the recovery of the remains of the extinct marine colonial organisms, graptolites, from Tremadoc rocks resulted in the inclusion of the Tremadoc in the Ordovician System. Planktonic graptolite remains do not occur below the Tremadoc Series.

The presence of graptolites in Ordovician strata led Lapworth and two of his students, Gertrude Elles and Ethel M.R. Wood, to propose graptolite zones as divisions of such strata. Lapworth formulated a graptolite zonal scheme in 1879. It was refined in 1918 by Elles and Wood, following extensive studies of British graptolite-bearing rock successions.

Establishment of the Lapworth and Elles–Wood graptolite zonal divisions of the Ordovician resulted from recognition that two distinct suites, or sets of strata, existed among Ordovician rocks. These two suites could be distinguished on the basis of the prominent fossil content of the rocks. One suite included strata containing the remains of graptolites. The second suite included rocks bearing the fossils of brachiopods, bryozoans, corals, and

Table 6: Selected Major Divisions of the Ordovician System and Their Correlation

British series	North American series	Australian stages	Chinese stages	Baltic region series
	Cincinnati			Harju
Ashgill		Bolindian	Wufengian	
			Linhsiangian	
		Eastonian		
			Pagodaian	
Caradoc	Mohawk			
		Gisbornian		
				Viru
			Huloian	
Llandeilo		Darriwillian		
			Niushangian	
Llanvirn		Yapeenian		
	Whiterock	Castlemainian	Chongyian	
		Chewtonian		
		Bendigonian	Ningkuoian	
Arenig				
	Ibex			
				Oeland
		Lancefieldian/ Warrendian		
Tremadoc			Xinchangian	

other organisms that secreted shells. The two basic suites of fossiliferous rock came to be called the graptolitic and the shelly-fossil facies. Because most graptolites are limited to the graptolitic facies and most shelly fossils occur only in the shelly facies, establishing synchroneities (*i.e.,* time correlations) between the graptolite zones and shelly-fossil divisions has been and continues to be an important area of study.

Economic significance. Ordovician rocks have proved commercially useful. Several varieties are used as building stone, in glass manufacture, and in cement making. In addition, certain dark shales are sources of petroleum, while some limestones, notably those in the Missouri River valley area of the United States, are hosts for lead and zinc ores.

Ordovician rocks. *Types and distribution.* Fossiliferous Ordovician strata occur on all present-day continents. In general, these rocks may be divided into shelly-fossil-bearing facies and graptolitic facies, as noted above. The shelly facies may be subdivided into two major rock suites. One of them, the carbonate suite, includes limestones and dolomites formed in supratidal to moderate-depth shelf-sea environments. Such environments and sediment accumulations occur most commonly in the latitudinal range of 30°–35° S to 30°–35° N latitude. The carbonate rock suite formed in shelf-sea environments on lithospheric plates that were within the tropics during the Ordovician. The second shelly-fossil-bearing rock suite consists primarily of siliclastic rocks, including various types of sandstones and siltstones and some mud rocks. The siliclastic shelly-fossil rock suite formed in marine shelf environments around Gondwana in nontropical latitudes and in environments where plate motion led to the development of a landmass. Such a landmass was formed along what is today the Appalachian Mountain region in eastern North America and the site of the ancient Caledonian orogenic belt that extended from Ireland and Scotland northeastward through Scandinavia.

Major rock suites

Thinly laminated, nonbioturbated (sediments undisturbed by organisms), black, graptolite-bearing shales are found rimming many successions of siliclastic rock containing shelly fossils. Similarly, thinly laminated, black, nonbioturbated graptolitic limestones rim shelly-fossil-bearing carbonates. The rocks with graptolites accumulated in marine environments that ranged from the deeper parts of the continental shelf, slope, and rise to parts of the open sea. The source of the graptolite-bearing rocks was primarily the materials accumulating in the shallow shelf and intertidal environments.

Several plates—including Laurentia, Siberia, Kazakhstania, North China, part of Southeast Asia, Australia, and part of Antarctica—that were aligned in the tropics during the Ordovician (Figure 20) are sites of Ordovician carbonate rock suites. Dolomites and evaporites (in the main gypsum and anhydrite) commonly occur in the central parts of these plates. Limestones rim the dolomite and gypsum–anhydrite–salt sequences. The dolomite–gypsum–anhydrite–salt sequences appear to have formed in the shallowest tropical shelf-sea environments. Since these sequences may be more than 2,000 metres thick and may have accumulated during only a part of the Ordovician, the plates are thought to have been subsiding throughout the period. Whether or not this plate subsidence is related to some form of lateral plate motion has not been documented. The movement of the Balto–Scanian plate into the tropics during the Ordovician was accompanied by the accumulation of carbonates in shelf-sea environments on that plate as it moved northward.

Siliclastic shelly-fossil rocks occur in western South America, North Africa, the Middle East, southern Europe and the British Isles, and parts of southern China. Glaciomarine sequences are common in the Late Ordovician (Late Caradoc and Ashgill) parts of the rock sequence in these areas. The Late Ordovician segment of the rock succession in central northern Africa contains glacial materials, including moraines and kame terrace deposits.

Correlation. Correlations among the Ordovician facies and rock suites are difficult, because graptolites rarely occur with shelly fossils and shelly-fossil-bearing rocks are seldom found in rocks with graptolitic facies. Most of the synchroneities that have been established between the facies are achieved by finding rock debris that contains shelly fossils or gravity-flow materials in graptolitic sequences. Because the several carbonate suite-bearing plates appear to have been separated longitudinally, synchroneities among rock suites on each plate have proved to be difficult to achieve. A set of Ordovician divisions is recognized for application in rock sequences on each plate. Correlations among these divisions have not been established definitively, however. Certain correlations among rock suites are shown in Table 6.

Gondwana

Ordovician environment. *Paleogeography and paleoclimate.* During the Ordovician, the major landmass, Gondwana, extended from the South Pole northward into the tropics (see Figure 20). The geologic history of areas included in Gondwana indicates that the continent fragmented during post-Ordovician time (see below *Mesozoic Era*). A number of major lithospheric plates developed from those that made up Ordovician Gondwana. Moreover, several large plates already existed within the tropics during the period. Most of them appear to have been centred approximately on the equator. As can be seen from the map, the Northern Hemisphere north of the tropics consisted primarily of open ocean. The plate on which modern Scandinavia and the Baltic area are situated moved from south of the tropics into the tropics during the Ordovician. Continental glaciers formed in the area of modern central and northern Africa. This area was close to the South Pole during the Late Ordovician. Pronounced glacio-eustatic lowering of sea level, perhaps as much as 100 metres, took place as a consequence of the South Polar glaciation. Late Ordovician glaciation persisted for about 8 to 10 million years.

Significant worldwide marine regressions occurred near the end of the Tremadocian and Llanvirnian ages. These regressions appear to have resulted from plate motion rather than from glaciation.

Ordovician environmental controls on organismal distribution. Geochemical and fossil analyses of Ordovician rocks suggest that the carbonate suite formed in shelf-sea environments analogous to those of the modern Florida Keys and Bahama Banks. The fossil content of the dolomites and associated limestones is primarily algae (notably the mat-forming variety) and nautiloids and small snails. The limestones bear the remains of brachiopods, bryozoans, crinoids, tetracorals, tabulate corals, and sponges and spongelike organisms. Trilobite fossils are not as common in the carbonate rocks as they are in the siliclastic sequences. The latter, on the other hand, contain significantly fewer brachiopods, bryozoans, corals, and crinoids than do the limestones. Besides the remains of trilobites, diverse trace fossils are found more commonly in the siliclastics.

Most siliclastic successions formed in Southern Hemispheric, nontropical marine shelf environments in which surface waters were cooler than in tropical shelf seas. Source terrains may have been relatively nearby; thus many of the shallow-shelf environments could have been sites of (at least seasonally) turbid waters. Graptolitic sequences appear to have formed in marine shelf, slope, and open ocean environments in which waters beneath the surface had little to no oxygen. Geochemical studies of Ordovician rocks suggest that the oxygen content of the atmosphere and, accordingly, of the surface ocean water was less than it is today. If this were indeed the case, then most marine environments would have had less dissolved oxygen than analogous modern environments. Because there is less dissolved oxygen in warm waters than in cool waters, the tropical shelf seas could have contained very little oxygen close to the water–sediment interface. Nontropical shelf seas could have been mixed by storms, perhaps seasonally. In addition, because the oxygen content of these shelf seas was greater than that of the tropical shelf seas, relatively more oxygen was available for benthic organisms living in marine shelf-sea environments in nontropical latitudes.

Effects of low oxygen content of oceanic waters

During the Ordovician, waters analogous to those of the modern oceanic-oxygen minimum zone may have been anoxic. Furthermore, such oxygen-depleted waters probably extended markedly deeper in the Ordovician oceans than they do in modern oceans. Certain Ordovician deep-seafloor sediments, notably red mudstones and cherts found in western Newfoundland, suggest that the deep oceans were at least moderately ventilated.

Effect of plate positions on oceanic circulation

Ordovician oceanic circulation. Plate positions based on the analyses of remanent magnetism, faunal provinces, and rock facies have implications for oceanic surface-water circulation. The absence of major plates in the Northern Hemisphere north of the tropics suggests zonal surface-water circulation above 30° N latitude. Insolation of the Earth's surface indicates that polar waters (those from the pole to 60° latitude) were colder than those south or north of 60° latitude. Accordingly, the surface waters between 60° and 90° N flowed east to west, and those between 30° and 60° N flowed west to east, in general. Tropical water flow across and around the plates in the tropics resulted in western boundary currents on the west side of Laurentia, Siberia, and the Middle East–South China region during much of the Ordovician Period. Oceanic surface circulation in the Ordovician Southern Hemisphere was influenced by the positions of plates bearing lands and shallow shelf seas. In addition, closure between the Balto-Scanian plate and Laurentia altered surface circulation between the two during the Ordovician. A western boundary current probably flowed northward along the Balto-Scanian plate during much of the period.

The presence of the major mass of Gondwana in tropical as well as in temperate latitudes suggests that monsoonal seasonal conditions could have developed near and over it. If this were so, then, by analogy with the surface circulation of the modern Indian Ocean, circulation of surface waters on the western side of Gondwana could have reversed seasonally. The mixing of tropical- and temperate-water planktonic faunas (graptolites) in rock sequences in southern China implies that such a surface circulation reflective of seasonal monsoon conditions could have taken place. Moreover, the richness of these faunas indicates that the western boundary currents along the Middle East–South China coasts created upwelling conditions for most of the Ordovician.

Ordovician life. The close of the Cambrian Period was marked by a mass mortality among the trilobites. As they had been the predominant shelf-sea marine invertebrate, the sharp reduction in their numbers opened many marine environments to colonization by other animal forms. Orthid brachiopods radiated (*i.e.,* rapidly dispersed into different environments) noticeably during the Ordovician. Strophomenid and rhynchonelloid brachiopods appeared and moved into many shelf-sea environments. Bryozoans, crinoids, and both tabulates and tetracorals appeared for the first time. Their appearance and initial radiation was in Ordovician tropical shelf-sea environments. Trilobites reradiated, particularly in the siliclastic, nontropical shelf seas. Bivalve mollusks diversified modestly in nearshore siliclastic environments. Nautiloids radiated significantly in the shallow, algae-rich tropical shelf seas early in the Ordovician. During that time, many nautiloids evolved to nektonic modes of life, possibly as a result of the ability to generate gas in their shells. Snails and monoplacophorans lived in the same shallow-shelf, carbonate environments as the nautiloids.

Graptolites with planktonic modes of life appeared during the Tremadocian Age. During the remainder of the Ordovician, planktonic graptolites underwent a major transformation, changing from colonies with many branches to those with only a few. In connection with this development, graptolite colonies that formed from two branches appeared. In the early varieties of such graptolite

Adapted from: C R Scotese, The University of Texas at Arlington

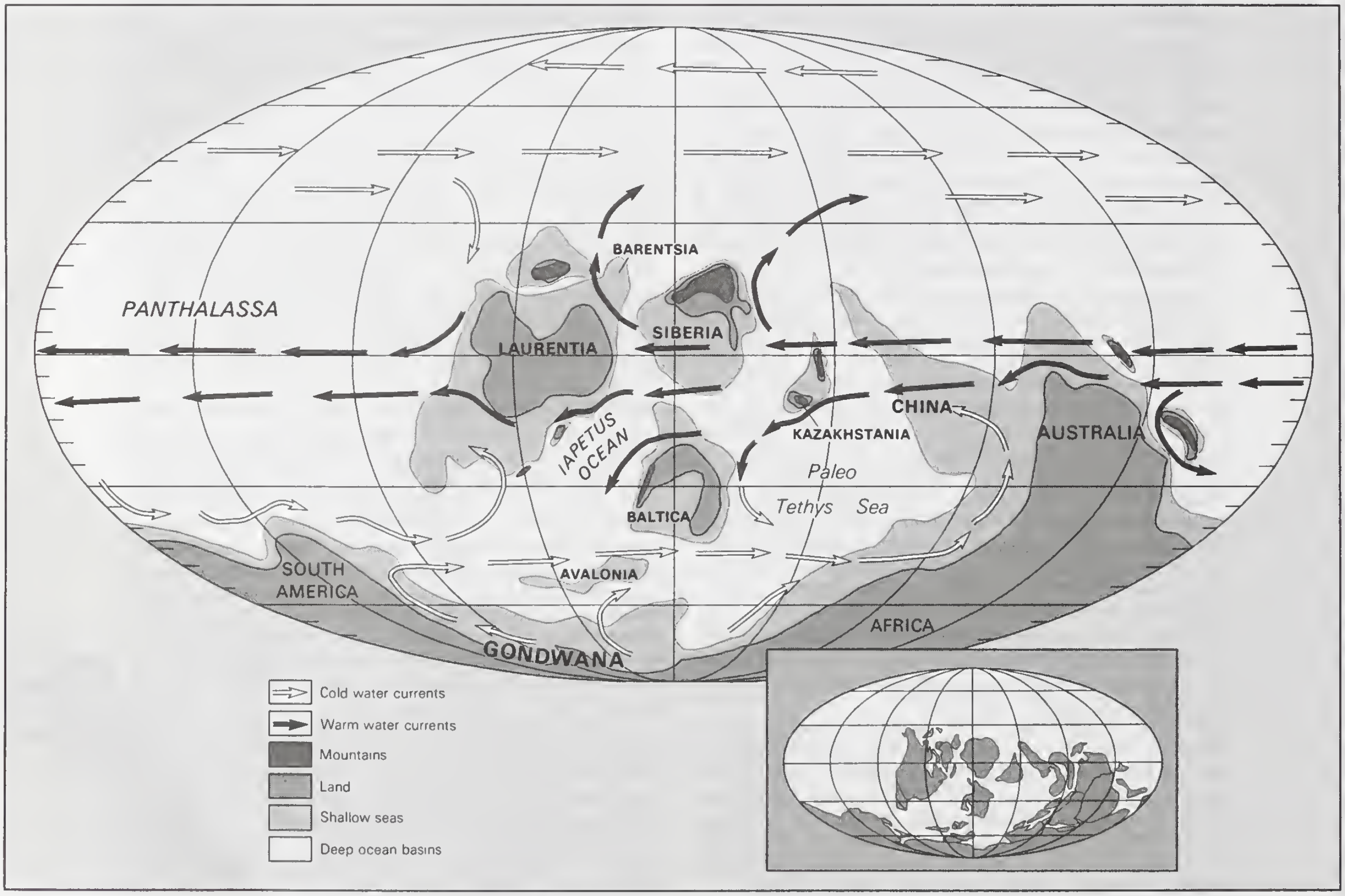


Figure 20: Distribution of landmasses, mountainous regions, shallow seas, and deep ocean basins during the Middle to Late Ordovician. Included in the paleogeographic reconstruction are cold and warm ocean currents. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset at the lower right.

colonies, the two branches were oriented in such a way that the zooids along them faced downward from the sea surface. In the later varieties, the zooids faced toward the sea surface. Zooids in the many-branched colonies faced downward from the sea surface.

The geologically oldest remains of nearly complete fish skeletons occur in Early Ordovician strata in Australia. These fish fossils are imprints on sandstones formed in nearshore marine environments. The remains are of ostracoderms, jawless armoured fish. Other, but disarticulated, ostracoderm remains are known from rocks of slightly younger Ordovician age in the western United States. The Ordovician ostracoderms occur in nearshore tropical marine settings.

Impact of prolonged continental glaciation

All marine animals suffered mass mortalities during the Late Ordovician. Presumably, these large-scale die-offs, observed among both shelly-fossil organisms and graptolites, were the result of environmental changes associated with the prolonged continental glaciation in the Southern Hemisphere. Mass mortalities at the close of the Cambrian and late in the Ordovician resulted in the unique aspects of the Ordovician fauna. The Late Ordovician mortalities created new opportunities for benthic and planktonic marine organisms. Reradiation during post-Ordovician glaciation led to many new taxa—those characteristic of the Silurian (see below *Silurian life*).

Some form of terrestrial plant life may have developed during the mid-Ordovician. Spores suggestive of land, but not vascular, plants occur in siliclastic Ordovician rocks on those lithospheric plates that were in the tropics at the time.

Distribution patterns of Ordovician marine organisms are consistent with plate distributions deduced from studies of remanent magnetism. The Ordovician tropical shelf seas were sites of a warmwater fauna distinct from that found in the nontropical shelf-sea rock suites. Studies of both benthic and planktonic organisms suggest that the several tropical plates were separated from one another. Faunas suggestive of distinct provinces may be found in rock suites on each plate. Faunal provincialism was marked during the Early Ordovician. As the plate bearing what is now the Baltic region and Scandinavia moved into the tropics during the mid-Ordovician, many new taxa were introduced to tropical environments. A marked incursion of such new taxa, seen in both planktonic and benthic faunas, typifies the basal part of the Caradoc Series. The ancestors of some of these taxa are found in the Balto-Scanian rock suites of Early Ordovician age. Ordovician faunal distribution patterns appear to reflect plate positions and relative plate motions, documenting the influence of plate tectonics on the distributions of organisms. (W.B.N.B.)

SILURIAN PERIOD

General considerations. The Silurian, the third period of the Paleozoic Era, occurred between 438 and 408 million years ago (see Table 4). As previously explained, absolute figures like these are calculated on the basis of radioactive isotopes, which must be analyzed geochemically from igneous samples. The radioisotopes decay at an exponential rate starting with the crystallization of the host rock from a magmatic source. Reheating of the host rock due to metamorphism, however, has the effect of resetting the radioisotope clock. Igneous rocks associated with sedimentary layers bearing fossils defined as Silurian are the desired targets for geochemical analysis, but few have remained untouched by metamorphism during their long existence. Typically, no more than four or five igneous localities are selected by teams of geochronologists to establish Silurian dates. Because igneous bodies are seldom found near the defined boundaries between geologic systems, the boundary ages for the Silurian Period must be extrapolated, taking into account a limited number of data points from older Ordovician and younger Devonian rocks. Competing groups of geochronologists disagree on precisely when and for how long the Silurian Period occurred. The oldest suggested age is 445 million years; the youngest is 395 million years. Some researchers believe the Silurian was as brief as 18 million years, while others argue for a span of 40 million years. In any case, the Silurian qualifies as one of the shortest geologic time periods, with many others approximately twice as long.

In 1835 R.I. Murchison named a sequence of rocks in Wales and its borderland with England "Silurian" in honour of a native people called the Silures, who had resisted Roman conquest. Murchison's classic contribution, *The Silurian System* (1839), illustrated 656 fossils, most of which were defined as characteristic of Silurian time. In this way, the formal groundwork was laid for recognition of Silurian rocks elsewhere around the world. Some years earlier (1822), William Buckland of Oxford University observed that fossil brachiopods and corals collected by British army officers stationed on Drummond Island (Michigan Upper Peninsula) and by fur traders working at Cumberland House (east-central Saskatchewan) in North America were equivalent to fossils from Dudley, Eng., as well as to those from the Swedish island of Gotland and the Baltic countryside (Estonia) neighbouring St. Petersburg. Each of these sites rests on Silurian bedrock, as now strictly delimited. Murchison's original concept grew in his lifetime to embrace what is now differentiated into the Cambrian, Ordovician, and Silurian systems. Claims for a monolithic system began to weaken in 1854 with discovery of the famous Onny River unconformity in Shropshire, Eng., which indicates a natural break within the classic Silurian on its own home territory. Major unconformities exhibiting a sustained pause in sedimentation subsequently were recognized between the Ordovician and Silurian systems in many other places around the world where the boundary beds are exposed.

Murchison's genius for military-style organization promoted the rapid adoption of his Silurian System throughout the British Isles, Scandinavia, and Europe. He was an active traveler who conceived of his field studies in terms of "campaigns." Local guides and fossil collectors were considered "aides-de-camp" in his service. Contemporaries who undertook research on Silurian strata and fossils with equal vigour were the French paleontologist Joachim Barrande in the Prague Basin and the American geologist and paleontologist James Hall in the states of New York, Michigan, Wisconsin, and Iowa. Other investigators who circulated to the distant territories of the British empire lost little time in recording the worldwide distribution of Silurian fossils. The first report of Silurian fossils in Australia, for example, was made (erroneously) in 1838, prior to the publication of *The Silurian System.* While the absolute limits of Silurian time remained imprecise, Murchison's triumph was the identification of successive fossil types with broad geographic distributions that always follow in the same relative stratigraphic order wherever they occur.

Boundaries and subdivisions. Continuing research has been preoccupied with decisions as to exactly where Silurian life may be defined stratigraphically as starting and ending and how the Silurian System may be broken into smaller chronostratigraphic units. The International Commission on Stratigraphy and its committees or working groups are charged by the International Union of Geological Sciences (IUGS) to reach such decisions based on the informed knowledge of members with broad geographic representation. The results of lengthy deliberations by the working group on the Silurian-Devonian Boundary were published in 1977, fixing the base of the *Monograptus uniformis* biozone (a single graptolite taxon but reinforced by associated conodont and trilobite taxa) as the base of the Devonian System. The top of the Silurian System is constrained by this marker (the so-called golden spike), which has its stratotype at a designated horizon in a cliff section near Klonk in the Czech Republic. Thus, the Silurian-Devonian boundary is anchored to the first occurrence of specific index fossils. The Klonk section acts as a kind of standard reference section with which other stratigraphic sections, potentially involving the Silurian-Devonian boundary beds, may be compared. This agreement arrived at by a committee of specialists represents the first time that the concept of the golden spike was put into effect internationally.

Boundary stratotype

In 1985 the working group on the Ordovician-Silurian Boundary ratified its decision to use the base of the

Parakidograptus acuminatus biozone (a group of concurrent graptolites) as the base of the Silurian System. The stratotype was fixed at a horizon in Dob's Linn near Moffat in the Southern Uplands of Scotland. The golden spike marking the Silurian–Devonian boundary is not associated with any unusual climatic changes or other physical phenomena that might have left a stratigraphic signature. In the case of the Ordovician–Silurian boundary, however, the effect on sea level of Late Ordovician glaciation and deglaciation with increasing deglaciation during the early Silurian accounts for widespread stratigraphic unconformities that usually omit the *P. acuminatus* biozone. In earliest Silurian time, the Dob's Linn locality was situated environmentally in marine waters deep enough to remain unaffected by these changes.

Silurian series

The IUGS's Subcommission on Silurian Stratigraphy has exercised its authority to codify chronostratigraphic subunits within the Silurian System. These include definitions of four series (see Table 8). The Llandovery, Wenlock, and Ludlow series bear names corresponding to historical units originally proposed by Murchison, but these are now rigorously defined in terms of basal stratotypes. The Wenlock Series, for example, derives its name from the Wenlock Edge (a prominent 25-kilometre-long ridge of limestone running southwestward from near the town of Much Wenlock in Shropshire). Its basal stratotype is now fixed at a locality called The Leasows section in Hughley Brook. The occurrence of specific microfossils (conodonts and acritarchs) signifies the beginning of Wenlock life and time as recorded at the base of the Wenlock series. Wenlock life and time may be said to reach an end when Ludlow life takes over (also represented by specific index fossils in a fixed stratotype). The last of the four series, the Přídolí, takes its name from an area outside Prague where a basal stratotype is defined by the first occurrence of a graptolite species (*Monograptus parultimus*) fixed at a horizon in exposures leading to the Pozary Quarries.

Silurian stages

Just as the Silurian System is composed of all rocks formed or deposited during the Silurian Period, the various series include all rocks formed or deposited during their respective time units called epochs. Series may be further divided into stages. The Llandovery Series consists of three stages (Table 7), whose names are derived from place-names for Welsh farms. The Wenlock and Ludlow series are each divided into two stages, but the Přídolí awaits further subdivision. All are defined by basal stratotypes, some of which are shared in common with larger units. The basal stratotype for the Rhuddanian Stage, for example, is the same as for the Llandovery Series and the Silurian System itself (located at Dob's Linn as discussed above). The various stages include all rocks formed or deposited during their respective time units called ages.

Table 7: Chronostratigraphic Units of the Silurian System

chronostratigraphy global standard stratigraphy				location of basal boundary stratotype
Silurian System	Upper Silurian	Přídolí Series	(division into stages to await necessity)	Barrandian (Pozary Section)
		Ludlow Series	Ludfordian Stage	Ludlow District (Sunnyhill Quarry)
			Gorstian Stage	Ludlow District (Pitch Coppice)
	Lower Silurian	Wenlock Series	Homerian Stage	Wenlock District (Whitwell Coppice)
			Sheinwoodian Stage	Wenlock District (Hughley Brook)
		Llandovery Series	Telychian Stage	Llandovery District (Cefn Cerig Section)
			Aeronian Stage	Llandovery District (Cefn Coed-Aeron Farm)
			Rhuddanian Stage	Southern Uplands of Scotland (Dob's Linn)

By permission of the National Museum of Wales

Regardless of their lithology, exposed fossiliferous strata, for instance, in the upper Mississippi River valley of the United States, may be classified as belonging to the Silurian System, the Llandovery Series, or the Telychian Stage, so long as they agree with fossils known to be restricted to those units as demarcated by their respective stratotypes in the United Kingdom and the Czech Republic. Correspondingly, the strata bearing particular index fossils in the upper Mississippi River valley may be described as having accumulated during the Silurian Period, Llandovery Epoch, or the Telychian Age.

Distinctive features. This hierarchy of stratigraphic nomenclature does little justice to the wealth of natural scenery shown by Silurian formations widely scattered around the world. Niagara Falls and the 11-kilometre-long Niagara Gorge on the Canadian–U.S. border were eroded and continue to be sculpted by rushing waters undercutting the soft Rochester Shale beneath a ledge of more resistant Lockport Dolomite. The Niagaran Escarpment is an arcuate ridge of resistant dolomite stretching more than 1,000 kilometres from the Niagara Falls area through the Bruce Peninsula and across Manitoulin Island (fringing the eastern and northern sides of Lake Huron in Ontario) to the Michigan Upper Peninsula and beyond to Wisconsin's Door Peninsula (fringing the northern and western sides of Lake Michigan). This resistant feature stands as much as 125 metres above the Great Lakes, which were shaped by the excavation of soft shales during the glaciations of the Pleistocene (see below *Pleistocene Epoch*).

Other notable manifestations of Silurian rock include the rolling hills of eastern Iowa and the similar rounded hills, called klintar, that dot the island landscape of Gotland, Swed., where Silurian mound reefs reach the surface. The renowned naturalist Linnaeus sketched in his field notebook the bizarre shapes of "stone giants"—large limestone sea stacks—8 to 10 metres high, which still stand in ranks along the shores of Gotland at Kyllej.

Some of Norway's beautiful inland (or isolated) fjords, such as Tyrifjorden northwest of Oslo, are lined by Silurian shales and limestones. Long graceful curves made by the Dniester River in Ukraine and the Moiero (or Moyyero) River in Siberia carve through high bluffs of Silurian limestone and marl. Picturesque sea cliffs formed by Silurian clastics guard the coasts of Ireland's Dingle Peninsula. Australia's Kalbarri National Park features river gorges winding their way to bold sea cliffs on the Indian Ocean, which are all set in Silurian Tumblagooda Sandstone. The partly Silurian Tabuk Formation forms vast desert stretches in Saudi Arabia. At an elevation of 6,000 metres, the Spiti River valley in India's Himalayan region is lined partly by limestone and quartzite belonging to the Muth Formation. Beneath the present varied scenery lies a totally different Silurian world, one which must be reconstructed from environmental clues in the rocks.

Elevated sea level

The most unusual features of the Silurian that distinguish it from the present-day physical environment relate to conditions of low continental elevations in conjunction with a much higher global stand in sea level. During Llandovery and Wenlock times, more than 65 percent of North America's craton (a somewhat smaller continent named Laurentia) was flooded by shallow seas ranging in water depth from a few to little more than 100 metres. These seas were tropical to subtropical in climate, and coral mound reefs with associated carbonate sediments were very common. During Ludlow and Přídolí times, deposition of evaporites (salts) was periodically set in motion as a result of reduced circulation. This general scenario had its counterparts in northern Europe (a separate continent named Baltica) and in Siberia (also a separate continent).

To a lesser degree, some of these elements are found in South China on the Yangtze platform (a separate continent) and Australia (part of the much larger continent Gondwana).

Seafloor topography was muted over large areas of these particular platforms, and faunas of shelly invertebrates were remarkably consistent with one another. Eastern Australia sustained a more varied seafloor topography due to extensive volcanism but shared many of the same faunal elements because of its tropical latitude. Much of remaining Gondwana (including Antarctica, India, Arabia, Africa, and South America) was centred over the south geographic pole, where a substantial landmass was heavily glaciated during late Ordovician time but much less so during early to mid-Silurian time. Early Silurian marine faunas recovered from a major extinction brought on by climatic change and lowered sea level resulting from late Ordovician glaciation. Except for primitive vascular plants restricted to coastal areas, the relatively small land areas were essentially barren.

Economic significance. Silurian rocks embrace a normal range of types that assert some role on local economies. A minor amount of petroleum is associated with Silurian reef structures; a substantial quantity of Silurian salt is mined; Silurian limestone and dolomite are widely quarried for crushed rock. The economic significance of Silurian raw materials, however, is mostly of historical relevance. Development of the Severn River valley in Shropshire, Eng., brought together for the first time the mineral ore, coal, and limestone necessary to fuel industrial iron production. The construction in 1779 of an iron bridge across the Severn River may be regarded as the starting point of the Industrial Revolution. Local quarrying of the Wenlock Limestone provided the fluxing agent necessary for the manufacture of iron. The English iron industry later shifted to the Birmingham area, where the Wenlock Limestone continued to be exploited for this purpose. A major underground canal system was built at Dudley in order to facilitate limestone mining. A similar juxtaposition of raw materials led to the industrial development of Birmingham, Ala., in the southeastern United States. Again, Silurian rocks provided one of the key ingredients. This time it was hematite ore from the Llandovery Red Mountain Formation, which was mined from 1862 to 1971. A third unusual site in this regard is the ghost town of Fayette in Michigan's Upper Peninsula. It was founded as a company town in 1867 because local resources offered an abundance of Silurian dolomite for use in iron smelting. At the opposite end of the Michigan Upper Peninsula on Drummond Island, dolomite from the Wenlock Engadine Group is still quarried on a large scale for this special industrial use.

Sedimentary rocks

Silurian rocks. *Types and distribution.* Excluding peat and coal, the same kinds of strata in the process of forming today were also deposited during Silurian time. Owing to the heightened state of sea level, coupled with the low relief of many continents, production of certain Silurian sediments was proportionately different than that observed in the present world, however. Chief among these are limestones, which form primarily from the carbonate detritus of coral skeletons, shells, and calcified algae. Unless such detritus is produced in great quantities or rapidly buried, it tends to dissolve in cold (temperate to polar) waters. In shallow warm (tropical to subtropical) waters, carbonates may collect more gradually to form continuous layers of limestone. The geographic locations of Laurentia, Baltica, and in part Siberia within 30° latitude on either side of the Silurian equator ensured the development of extensive platform carbonates. In North America, Silurian limestones or dolomites (altered from limestone by partial secondary substitution of magnesium for calcium) are found across an enormous territory stretching along one axis from northern Greenland to West Texas and along another axis from Quebec's Anticosti Island to the Great Basin of Utah and Nevada. Formation names and their stratigraphic ranges are shown in Table 8 for some typical carbonate sequences preserved in Canada (Quebec and Manitoba) and in the United States (northern Michigan, eastern Iowa, and the Great Basin of Utah and Nevada). Parts of Baltica where carbonate deposition was prevalent include Sweden's Gotland, Estonia, and the Ukrainian region of Podolia; carbonate deposition was also prevalent over much of Siberia (Table 8). Platform carbonates of this kind rarely exceed 200–300 metres in thickness. Important limestone units more restricted in Silurian time and space include the Wenlock Limestone (Shropshire, Eng.), the Ryterraker Formation (southern Norway), the Xiangshuyuan Formation and lateral equivalents (South China), and the Hume Limestone (New South Wales, Australia).

Evaporites, including salt (halite), anhydrite, and gypsum, are chemical precipitates that usually accumulate as layers through evaporation of marine waters isolated in shallow bays. This process is most effective under a warm, arid climate commonly found at latitudes of about 30° or less. Distributed through parts of Michigan, Ohio, and New York state, the Upper Silurian (Ludlow–Pridoli) Salina Group is one of the world's most famous evaporite deposits. A maximum aggregate thickness of 600 metres occurs in Michigan, where one individual halite bed reaches a thickness of 165 metres. A 2-metre halite bed occurs in the Interlake Formation (Wenlock) of North Dakota. Gypsiferous beds occur in parts of the Upper Silurian Yangadin and Holuhan formations of Siberia, as well as in comparable formations in Latvia and Lithuania. Upper Silurian (Pridoli) evaporites are characteristic of three different basins in Western Australia. Minor amounts of halite and anhydrite occur in the Dirk Hartog Formation in the Carnarvon Basin; more extensive halite or anhydrite beds or those of both have been discovered in comparable formations from the Canning and Bonaparte Gulf basins.

Clastic wedges

Clastic rocks, including conglomerates, sandstones, and shales, generally occur in wedge-shaped deposits adjacent to land areas from which terrigenous materials erode under conditions of moderate to high annual rainfall. Owing to steady accumulation over protracted periods of time, such deposits tend to become very thick and subside under their own weight, forming troughlike structures parallel to their sediment source. In contrast to thin platform deposits, clastic wedges may be thousands of metres or more than one kilometre thick. Taconica was a long narrow highland roughly corresponding to the present position of the Appalachian Mountains in North America. During early Silurian (Llandovery) time, these highlands shed the Shawangunk Conglomerate (500 metres thick) near its front in southeastern New York state and distributed the Tuscarora–Clinch sandstones (150–250 metres thick) throughout central Pennsylvania and western Virginia from more than 80 kilometres beyond its front. These deposits accrued from sediments carried by braided streams crossing coastal plains to a wave-swept shore. *Arthrophycus* trails (those made by annelids tolerant of low salinity) are recorded in the more seaward portions of the Tuscarora Sandstone. Collectively attributed to the Clinton Group, a variety of Upper Llandovery rocks with high iron content subsequently were deposited from New York to Alabama. These strata often contain marine fossils, but their iron was derived from Taconica. Tiny pellets, or oolites, coated with hematite occur in seams up to 2 metres thick in New York; massive ferruginous sandstones are found in Pennsylvania; and oolitic ironstone beds up to 15 metres thick occur in Alabama (in the Red Mountain Formation).

Evidence of another Laurentian highland called Pearya is found in the Canadian Arctic in the vicinity of northern Ellesmere Island. Clastic sediments eroded from this source were deposited in the Hazen Trough. One Lower Silurian (Llandovery) unit called the Danish River Formation is composed of interstratified conglomerates, sandstones, and shales one kilometre thick. The Caledonian highlands dominated depositional patterns on the paleocontinent of Baltica. Much of the highland front followed approximately the present spine of Norway and affected a broader area through generation of river-transported sandstones that gradually spread across Sweden to Poland in one direction and through northern England to southeastern Ireland in the other direction. Known traditionally as the Old Red Sandstone, these rocks are of late Silurian (Ludlow) age in southern Norway, mixed late Silurian (Pridoli) and early Devonian age in northern England, and

Table 8: General Subdivisions and Regional Rock Sequences of the Silurian System

Lower Silurian				Upper Silurian		series	
Llandoverian			Wenlockian	Ludlovian	Pridolian	stage	
Lower	Middle	Upper					
shale, sandstone, and mudstone			Wenlock Shales Wenlock Limestone	Eltonian Bringewoodian Leintwardinian Whitcliffian	Ludlow Bone Bed lower half of Downtonian	Welsh borderland	Britain
Kilfillian Formation	?	? Garheugh Formation Hawick Rocks: Carghidown Beds ? Kirkmaiden Beds	Riccarton Beds ?			Southern Uplands	Britain
		unnamed Silurian strata (subsurface) Visby Marls	Högklint Group Slite Group Mulde Marl	Hemse Group Eke Group	Hamra Group Sundra Group	Gotland	Sweden
Rastrites Shale		Cyrtograptus Shale: ? Retiolites Beds	? Hemingü Beds	Colonus Shale ?	Oved-Ramsåsa Group	Skåne	Sweden
Liten Shales			Liten Limestone	Kopanina Limestone	Pridoli Limestone	Prague region	
Tabuk Formation						Arabia	
Argile Schisteuses						Morocco	Africa
Tanezzuff Shale				Aracus Sandstone		central Sahara	Africa
		Table Mountain Sandstone				South Africa	Africa
Juuru Stage Raikkula Stage		Adavere Stage	Jaani Stage Jaagarahu Stage	Kaarma Stage Paadla Stage Kuressaare Stage	Kaugatuma Stage Ohesaare Stage	Estonia	former U.S.S.R.
		Kitaigorod	Bagovista	Malinovetski	Skala	Podolia	former U.S.S.R.
	Kosyu Horizon	Adak Horizon	Filipelsk Horizon Sedelsk Horizon	Gerdyusk Horizon	Greben Horizon	western Urals (Chernisheyov Range)	former U.S.S.R.
	unnamed Llandovery Volcanics		unnamed Wenlock Volcanics	Elkino Series Zhuravlik Series	Turin Series	eastern Urals (Is River)	former U.S.S.R.
Moierokan	Haastyr	Agidyi	Hakom	Yangadin	Holuhan	Siberian Platform	former U.S.S.R.
Lungmachi ?	Xiangshuyuan	Leijiatun Rongxi Xiushan	Huixingshao ?			China (Yangtze region)	
Melbourne Group				Yering Group		Melbourne area	Australia
	? Mundoonan Sandstone ?	Hawkins Series ? Bango Group ? Douto Group ?	Yass Group ? Laidlaw Group ?	Barrandella Shale ? Hume Group: Hume Limestone ? Black Bog Shale	Dalmanites Beds ? green and black shale	Yass area	Australia
		? ? ?	Cancañiri Graywacke	Llallagua Formation Pampa Shale Catavi Sandstone ?	Ventilla Shale	Bolivia	
Caparo Formation				?		Venezuela	
Road River Group						Yukon	North America
Allen Bay Formation				Read Bay Formation		Arctic Canada	North America
Interlake Group (dolomites)						southern Manitoba	North America
		Laketown Dolomite		? Water Canyon		eastern Nevada and Utah	North America
Mosalem	Tête des Morts Blanding	Hopkinton Scotch Grove	Gower "Anamosa"			eastern Iowa	North America
Manitoulin Dolomite Cabot Head Shale Moss Lake Formation	Burnt Bluff Group: Lime Island Dolomite	Burnt Bluff Group: Byron Dolomite Hendricks Dolomite Manistique Group: Schoolcraft Dolomite, Cordell Dolomite	Engadine Dolomite	Point Aux Chenes Shale	St. Ignace Dolomite	northern Michigan	North America
Grimsby Sandstone		Kodak Sandstone Maplewood Formation Reynales Island Wallington Formation Upper Sodus Shale Williamson Shale Irondequoit Island: Rockway Member, upper member	Rochester Shale Lockport Group	Salina Group: Vernon Shale, Syracuse Salt, Camillus Shale	Bertie Group Akron Dolomite	New York State	North America
Ellis Bay Formation	Becsie Formation Gun River Formation	Jupiter Formation Chicotte Formation ?				Anticosti Island, Canada	North America
		Quoddy Formation ? Dennys Formation ?		Edmunds Formation	Pembroke Formation	southeastern Maine	North America
Carys Mills Formation Spragueville Formation				Perham Formation upper member		Presque Isle, Maine	North America
Glencoe Brook Formation	Kerrowgare Formation				Stonehouse Formation	northern Nova Scotia	North America
Beechhill Cove Formation lower member Ross Brook Formation	middle member Ross Brook Formation	upper member Ross Brook Formation	French River Formation Doctor's Brook Formation	McAdam Brook Formation Moydart Formation		northern Nova Scotia	North America

early Devonian age in southeastern Ireland and Poland. This variation in age reflects the growth of the Caledonian highlands and their ability to shed clastic debris farther and farther afield. In Western Australia, similar thick red sandstones belonging to the Upper Silurian Tumblagooda Sandstone were derived from a Precambrian massif called the Yilgarn Block. In contrast to sandstones that accumulated because of river transport, eolian sandstones are those deposited under desert conditions. The Mereenie Sandstone in central Australia (Amadeus Basin) is one of the few examples of a possible Silurian desert sandstone.

Outside clastic wedges closely linked to a land source, Silurian shales also formed on platform margins, as in the nearly 500 metres of strata belonging to the Road River Group in the Canadian Yukon. Based on sections in the Mackenzie Mountains, a distance of only one to a few kilometres separated the edge of a shallow water carbonate platform from the deepwater setting of basinal shales. Submarine avalanches (turbidity flows) brought the 1,200–1,500 metres of interbedded shales and fine sandstones constituting the Aberystwyth Grit Formation to a deepwater basinal setting in west-central Wales. Less commonly, Silurian shales passively accumulated in broad platform settings. The Lungmachi Formation in South China (Yangtze platform) is one such shale body, which makes up the basal Silurian throughout parts of the Yunnan, Szechwan, Shensi, Hupeh, Hunan, and Kweichow provinces. As much as 500 metres thick in places, these shales developed under anoxic conditions in quiet waters. Similar conditions prevailed during early Silurian times well within Baltica, including southern Sweden and Denmark.

Silurian sandstones and shales rest directly on Upper Ordovician tillites in Arabia (Tabuk Formation) and throughout large parts of North Africa. In South America (fused with Africa during the Silurian), glaciation persisted well into the period (Wenlock age). The Cancaniri Formation, including a prominent 60-metre-thick segment that bears the Zapla Tillite, extends 1,500 kilometres from northern Argentina over the Andes Mountains across Bolivia to Peru. Alpine glaciers descended to tidewater to deposit these layers, including faceted and glacially striated boulders 1.5 metres in diameter. Similarly, the widespread Trombetas Formation in the Amazon of Brazil may represent Silurian tillites filling submarine channels gouged by drifting ice blocks.

Volcanic rocks

Examples of rocks used to make absolute age determinations for the Silurian include a Llandovery volcanic breccia from the Descon Formation on Esquibel Island in Alaska, ash beds (bentonites) from the basal Wenlock Buildwas Formation and the Ludlow Elton Formation, both in Shropshire, Eng., and the Laidlaw Volcanics of Ludlow age near Canberra, Australia. Compared to other time periods, the Silurian was relatively quiet in terms of volcanic activity. Moderate activity occurred in those parts of the British Isles, the Canadian Maritimes, and coastal New England collectively attributed to Avalonia (as appended to Baltica). Approximately 1,000 metres of basalt flows belonging to the Skomer Volcanics in southwestern Wales are Llandovery in age. Rhyolite and andesite lavas were extruded in the area of the English Mendip Hills during Wenlock time. Basalts, rhyolites, and porphyritic andesites from the Newbery Volcanics in northeastern Massachusetts are Pridoli in age. Rhyolitic and andesitic flows of Silurian age also are known in the region of Passamaquoddy Bay in Maine, as are volcanic flows and breccias in adjacent New Brunswick. A Silurian chain of volcanic islands stood off the Laurentian craton, stretching from the Klamath Mountains of northern California to Alaska. Likewise, andesitic and basaltic sites of volcanism stretched along the edge of Baltica, as framed by the Ural Mountains.

The most extensive Silurian volcanism occurred in eastern Australia, principally in New South Wales but also in Victoria and Queensland. Activity was initiated during early Wenlock time with the Paddys River and Uriarra volcanics; expanded during Ludlow time with the Laidlaw, Mineral Hill, Bennetts Creek, Mullions Range, and Bells Creek volcanics; and concluded in Pridoli time with the Bombay Creek and Woodlawn volcanics in addition to the Currawan Basalt. Most of these were subaerial except for the basalt, which left submarine pillow structures. In New South Wales, more than 200 intrusive granitic bodies of the late Silurian or early Devonian are geochemically linked to these acid volcanics, but their precise age is difficult to establish.

Correlation. The most challenging goal in stratigraphy is to identify on a global basis all those rocks formed during the smallest possible interval of geologic time. Correlation of Silurian strata within limits more refined than a stage (or its corresponding age) is achieved through the recovery of fossils belonging to shaley and shelly facies.

Silurian index fossils

Shaley facies generally represent deeper-water environments, such as those under which the Road River Group (Yukon), Aberystwyth Grit Formation (Wales), and Lungmachi Formation (South China) accumulated. Fossils of graptolites are abundant in these dark Silurian shales. As explained above, graptolites were colonial hemichordates that secreted a protein exoskeleton commonly preserved as a carbon film in shales. An individual lived within a cuplike structure (theca); multiple cups were spaced along one or more branches (stipes); and the entire colony sometimes was connected by a threadlike structure (nema) to a central float. Some graptolites were bottom-dwellers, but the floating (or pelagic) species were geographically more cosmopolitan. They make excellent index fossils, because they underwent rapid evolution and attained a broad distribution. The genera *Pristiograptus* and *Cyrtograptus,* whose general features are illustrated in Figure 21, are pelagic graptolites characteristic of the Wenlock Series. As many as 42 graptolite biozones have been defined for the Silurian System (Table 9). Each biozone takes the name of one particular species but is usually based on several coeval species. The span of time represented by each graptolite biozone probably is not perfectly uniform, but the zonation facilitates correlation of strata in depositional units of 1 million years or less (provided of course that the Silurian Period lasted 30 million years). Superb as it is, this level of precision is restricted to regions rich in graptolitic shales. It is easier to correlate the deepwater shales of Wales and the Yukon with each other, for example, than it is to correlate either with nearby shallow-water shelf deposits.

In contrast to shaley facies, shelly facies are represented by relatively shallow platform carbonates and clastic wedges with a retinue of mostly bottom-dwelling invertebrates. Among these, Silurian brachiopods were especially

From E. Winson in W.S. McKerrow (ed), *The Ecology of Fossils*, Gerald Duckworth & Company Ltd

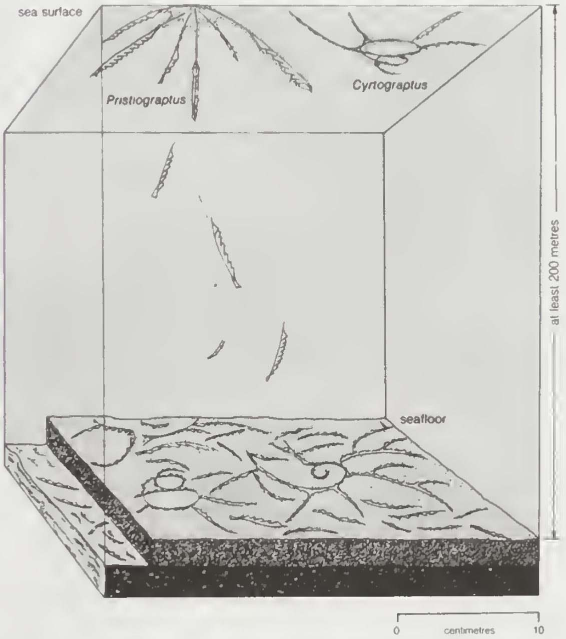


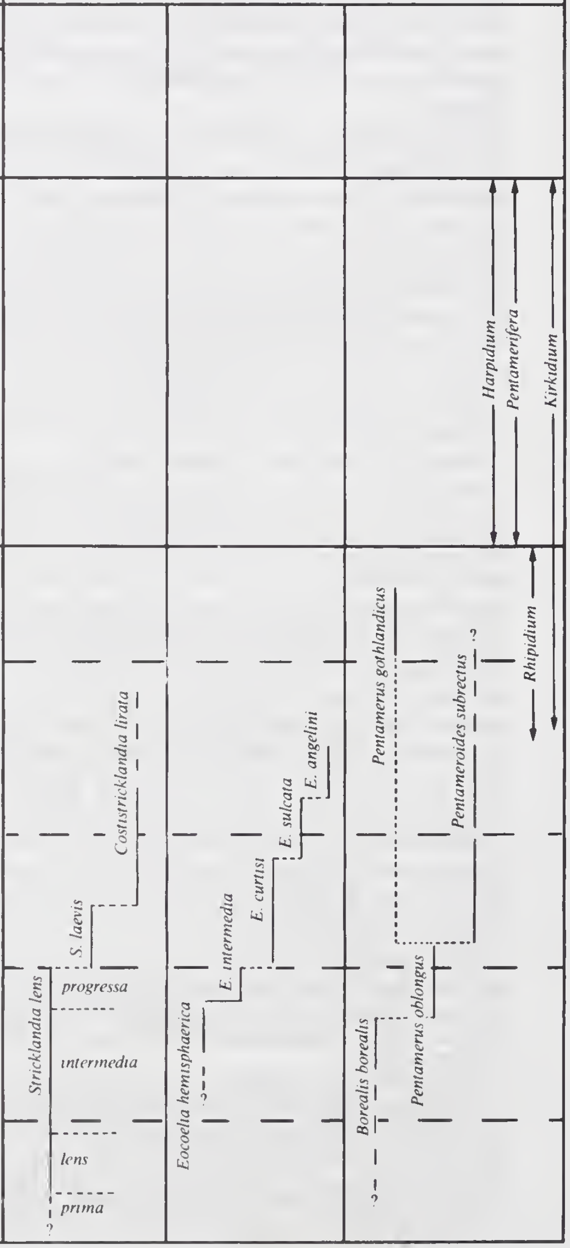
Figure 21: Pelagic graptolites belonging to the Wenlock genera *Pristiograptus* and *Cyrtograptus*.

abundant, diverse, and widely distributed, making them effective index fossils. A still extant group, the brachiopods possess a pair of bilaterally symmetrical shells and are tethered to the seafloor by a fleshy appendage usually protruding through one of the shells (thus the shells are typically unequal in shape). Biostratigraphic zonations based on brachiopod lineages are well suited to the correlation of Lower Silurian (Llandovery and Wenlock) strata. Those most frequently used are the *Stricklandia–Costistricklandia*, the *Borealis–Pentamerus–Pentameroides*, and the *Eocoelia* lineages (Table 9). Excluding most of Gondwana except eastern Australia, these brachiopods attained a broad tropical to subtropical distribution. Lineage members (including four subspecies of *Stricklandia*) may be employed independently but are more effective where it is possible to use them in combination as overlapping taxon zones. Thus, *Pentameroides subrectus* alone may indicate a late Llandovery (Telychian) age or an early Wenlock (Sheinwoodian) age. In association with *Stricklandia laevis* or *Eocoelia curtisi*, however, a Telychian age is certain. Single taxon zones also are helpful in the Upper Silurian (Ludlow and Pridoli). Only in unusual cases where brachipod-bearing layers are interbedded with graptolite-rich shales (at the narrow transition between shelly and shaley facies) may the temporal relationships between the two kinds of index fossils be established. Most of the projected correlations between graptolite and brachiopod biozones are approximate (Table 9).

Conodonts constitute a third group of index fossils important for Silurian correlation. These phosphatic microfossils with the shape of conelike teeth (as the name implies), are the remains of an apparatus from the mouth cavity of a small, bilaterally symmetrical, free-swimming (nektonic) animal extinct since Triassic time. Rare body fossils suggest some affinities with surviving cephalochordates, such as *Amphioxus*, or the chaetognaths (arrowworms). The individual elements comprising the conodont apparatus are very common in Silurian rocks that accumulated under a wide range of marine environments. Hydrochloric acid, which has no effect on phosphatic material, is used to dissolve limestone or lime-cemented sediments from which the conodonts may be recovered. Graptolite and brachiopod zonations have a long history of use in the Silurian, but the first conodont zonation based on material

Table 9: Ranges of Important Biozones in the Silurian System

series	stage	graptolite biozones	conodont biozones	brachipod biozones
Přídolí		*Monograptus transgrediens* *Monograptus perneri* *Monograptus bouceki* *Monograptus lochkovensis* *Monograptus pridoliensis* (= *M. simulis*) *Monograptus ultimus* *Monograptus parultimus*	*Icrioddus woschmidti woschmidti* *Ozarkodina remscheidensis eosteinhornensis*	
Ludlow	Ludfordian	*Monograptus balticus/caudatus* *Neocucullograptus kozlowskii* *Neocucullograptus inexpectatus* *Neolobograptus auriculatus* *Bohemograptus cornutus* *Bohemograptus praecornutus* } *Bohemograptus bohemicus* *Cucullograptus aversus* *Saetograptus leintwardinensis*	*Oarkodina crispa* *Ozarkodina snajdri* *Pterospathodus siluricus*	
	Gorstian	*Cucullograptus hemiaversus* } *Saetograptus incipiens* or *Pristiograptus tumescens* *Lobograptus invertus* *Lobograptus scanicus* } *nilssoni-scanicus* *Lobograptus progenitor* *Neodiversograptus nilssoni*	*Ancorodella ploeckensis*	
Wenlock	Homerian	*Pristiograptus? ludensis* *Gothograptus nassa* *Cyrtograptus lundgreni*	*Ozarkodina bohemica bohemica* *Ozarkodina sagitta sagitta* *Ozarkodina sagitta rhenana*	
	Sheinwoodian	*Cyrtograptus ellesae* *Monograptus flexilis* *Cyrtograptus rigidus* *Monograptus riccartonensis* *Cyrtograptus murchisoni* *Cyrtograptus centrifugus*		
Llandovery	Telychian	*Monoclimacis crenulata* *Monoclimacis griestoniensis* *Monograptus crispus* *Monograptus turriculatus*	*Pterospathodus amorphognathoides* *Pterospathodus celloni*	
	Aeronian	*Monograptus sedgwickii* *Monograptus convolutus* *Pribylograptus leptotheca* } *Coronograptus gregarius* *Diplograptus magnus* *Monograptus triangulatus*	*Disfomodus staurognathoides*	
	Rhuddanian	*Coronograptus cyphus* *Lagarograptus acinaces* *Atavograptus atavus* *Parakidograptus acuminatus*	*Disfomodus kentuckyensis*	



collected at Mount Cellon in the Carnic Alps of Austria was not proposed until the early 1960s. Different kinds of conodont-bearing animals lived in shallow, nearshore environments (as opposed to deeper, more offshore environments), and some were more conservative in their evolutionary development than others. Global correlation is based on $12\frac{1}{2}$ conodont biozones for the complete Silurian (Table 9). The base of each zone is defined by the first occurrence of the taxon. All are characteristic of open marine environments; the zonation is not applicable to strata that accumulated in restricted nearshore or deep-basin facies.

Global correlation based on conodont biozones

Silurian environment. *Paleogeography.* The Silurian world consisted generally of a north polar ocean, a ring of at least six equatorial to middle-latitude continents, and one south polar supercontinent. A combination of paleomagnetic, paleoclimatic, and biogeographic data can be used to reconstruct the approximate orientations of the Silurian continents. The Earth's magnetic field leaves its signature on volcanic rocks and certain sedimentary rocks rich in such minerals as magnetite. As rocks capable of being magnetized were cooled or otherwise lithified, their component crystals or grains conformed to an alignment with the terrestrial magnetic field consistent with their latitudinal origin. Unless the rocks were reheated or reworked by erosion, they should retain this signature regardless of subsequent geographic displacement. The Earth's zonal climate also has an effect on global patterns of sedimentation (see above). Strata formed in arid regions, for example, differ from those formed in regions with high annual rainfall. Endemism provides an index sensitive to continental isolation. The geographic summary that follows is based on a global reconstruction specific to middle Silurian (Wenlock) time (Figure 22).

Much of North America, including Greenland, northwestern Ireland, Scotland, and the Chukotsk peninsula of northeastern Russia, belonged to the paleocontinent Laurentia. (The name is derived from Quebec's portion of the Canadian Precambrian shield.) With respect to the present-day Great Lakes and Hudson Bay, Laurentia was rotated clockwise during Wenlock time to fully fit within latitudes 30° N and S of the paleoequator. The present south shore of Hudson Bay was at the centre of Laurentia, with the Wenlock paleoequator crossing near Southampton Island. The microcontinent Barentsia (including Norway's island of Svalbard) probably was appended to Laurentia off eastern Greenland. Island arcs and highland areas, such as Taconica and Pearya (see above), rimmed the flooded continent.

Paleo-continents

The narrow, north–south Iapetus Ocean still separated Laurentia from Baltica during Wenlock time. (The continent's name is derived from the Baltic Sea and Baltic states—Estonia, Latvia, and Lithuania—at the core of northern Europe.) The Uralian and Variscan–Hercynian sutures marked the eastern and southern margins of this paleocontinent, respectively. The northern tip of Scandinavia was situated just below the Wenlock paleoequator, but the islands of Novaya Zemlya extended well above it. The most prominent features were the Caledonian highlands of Norway, although a lowland may have existed in the vicinity of Finland. The microcontinent of Avalonia—its name derived from the Avalon Peninsula of eastern Newfoundland—probably was appended to Baltica by the end of Ordovician time. It included what is now England, Wales, southeastern Ireland, the Belgian Ardennes, northern France, eastern Newfoundland, part of Nova Scotia, southern New Brunswick, and coastal New England.

Separated from Baltica by the Pleionic Ocean, the paleocontinent of Siberia assumed an orientation rotated 180° from its present alignment (as recognized by the inverted

Adapted from C R Scotese, The University of Texas at Arlington

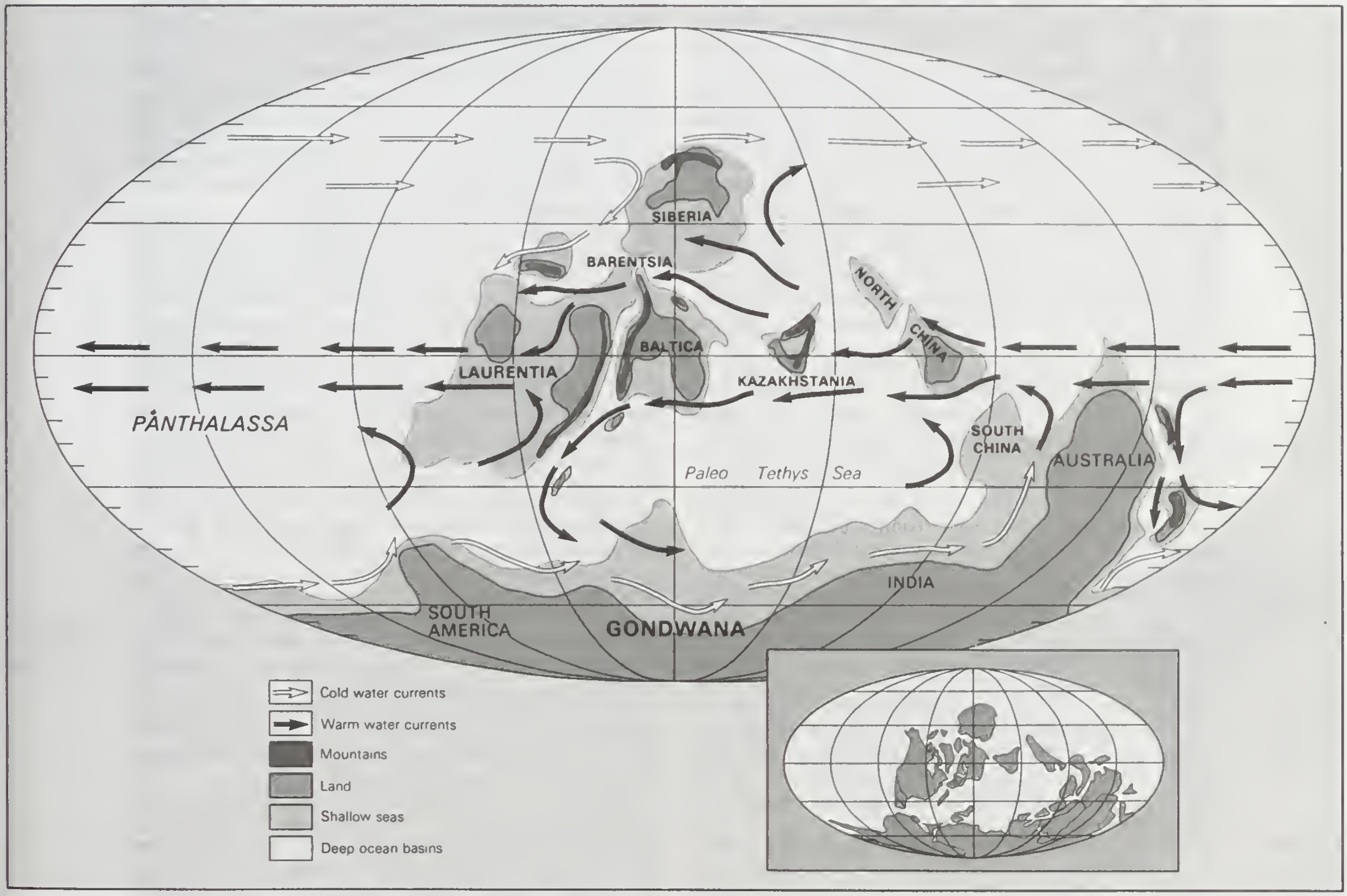


Figure 22: Distribution of landmasses, mountainous regions, shallow seas, and deep ocean basins during Early Silurian time. Included in the paleogeographic reconstruction are cold and warm ocean currents. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset at the lower right.

position of Lake Baikal). A huge Siberian platform sea was almost completely surrounded by lowlands. Kazakhstania was a neighbouring continent in the same middle northern latitudes. North China (including Manchuria and Korea) and South China (the Yangtze platform) were two separate continents situated in a more equatorial position. In contrast to the foregoing continents, most of North and South China were elevated above sea level during Wenlock time.

The vast supercontinent centred over the South Pole was Gondwana. In addition to Australia, Antarctica, India, Arabia, Africa, and South America, Silurian Gondwana also included smaller pieces of Florida, southern Europe, and the Cimmerian terranes—namely, Turkey, Iran, Afghanistan, Tibet, and the Malay Peninsula—on its outer fringes. The east–west ocean separating the southern European sector of Gondwana from northern Europe (Baltica) is called the Rheic Ocean. Present-day Brazil or contiguous West Africa was the locus of the South Pole, buried by an ice cap probably comparable in size to Antarctica. During Wenlock time, India, Tibet, the Malay Peninsula, and Australia projected into subtropical or tropical latitudes (see Figure 22).

Environmental conditions favourable to marine invertebrates

Significant geologic events. Resumption of environmental conditions favourable to faunal recovery from the late Ordovician extinctions was the most significant geologic event of the Silurian Period. It is estimated that these extinctions claimed 12 percent of all marine invertebrate families. Sixty percent of late Ordovician brachiopod genera survived the start of the Silurian Period, but only 20 out of 70 tabulate and heliolitoid coral genera and 14 out of 38 trilobite families made the same transition. Dramatic unconformities between the Silurian and Ordovician systems indicate how extreme the glacially induced drawdown in late Ordovician sea level had been. The maximum global fall in sea level was as much as 170 metres and drained immense areas of former marine habitat. River valleys up to 45 metres in relief were eroded into Upper Ordovician marine shales stretching across Iowa, Wisconsin, and Illinois on the Laurentian platform. On Baltica, marine carbonates in Norway and Sweden were transformed into karst surfaces through subareal exposure; a network of extensive tidal channels was developed across a formerly much deeper shelf in Wales. Close to the edge of the Gondwanan ice sheet in Saudi Arabia, the Jabal Sarah paleovalley was deeply incised by glacial outwash streams eroding through Ordovician shales and sandstones. These features and many others like them elsewhere were eventually filled and buried with the return of marine sedimentation in early Silurian time. Basal Silurian strata virtually everywhere record a rapid rise in the level of the sea, which reflooded vast continental platforms.

The few localities where marine sedimentation continued uninterrupted from late Ordovician to early Silurian time have been scrutinized for unusual trace elements. A major iridium spike of the sort widely found at the Cretaceous–Tertiary boundary has not been detected at the basal Silurian stratotype in Scotland nor at the parastratotype on Anticosti Island in Quebec. The introduction of high iridium levels in the environment requires unusual volcanic activity or the impact of a large asteroid. Neither of these agents seems to be linked with the Late Ordovician extinctions or subsequent Silurian developments. The small Lac Couture crater (eight kilometres in diameter) in Quebec is the only known impact structure of Silurian age, but no known iridium layer is associated with it. Large-scale reduction of marine habitat space in combination with global cooling probably were the primary agents of the Late Ordovician extinctions. Some Silurian continents recovered more rapidly than others. The niche of the large-shelled pentamerid and stricklandiid brachiopods on the Laurentian platform, for example, was filled belatedly by immigrants from Siberia and Baltica.

Recurrent fluctuations in sea level

Smaller fluctuations in sea level between 30 and 50 metres in magnitude continued to occur on a global basis throughout the Silurian. In contrast to the Late Ordovician event, these fluctuations did not strongly affect the shelly bottom-dwelling invertebrates perched on continental platforms. Benthic faunas accommodated by living conditions at particular bathymetric levels simply tracked changing sea level by sifting upslope or downslope. The amount of available habitat space was not drastically altered as a result of these maneuvers. Data from three or more different paleocontinents indicate that at least four globally coordinated highstands (*i.e.,* intervals during cycles of relative sea-level changes when the sea level lies above the continental shelf edge of a given area) took place during Llandovery time. The first event probably corresponds to the maximum rise in sea level achieved following recovery from the major drawdown in Late Ordovician sea level. This highstand occurred during the transition between the Rhuddanian and Aeronian ages near the boundary between the *Coronograptus cyphus* and *C. gregarius* graptolite biozones (Table 9). A second highstand of mid-Aeronian age corresponds to the basal parts of the *Stricklandia lens progressa* lineage zone and the *Monograptus sedgwickii* graptolite biozone. The third matches an early Telychian event linked to the *Stricklandia laevis* and *M. turriculatus* zones, and the fourth, a late Telychian event, is correlated with the lower range of the *Costistricklandia lirata* and *M. crenulata* zones. Standard geochronology suggests that these cycles of rising and falling sea level had an average duration of about 2.5 million years during the Llandovery.

Present data are not as complete for the rest of the Silurian, but a mid-Wenlock highstand in sea level is widely reported as coeval with the *Monograptus riccartonensis* to *Cyrtograptus ellesae* graptolite biozones. A mid-Ludlow lowstand in sea level also is commonly equated approximately with the *Saetograptus leintwardinensis* biozone, separating an early Ludlow highstand from at least one subsequent Ludlow highstand. Information on sea-level changes during Pridoli time is fragmentary and globally inconsistent. Late Silurian lowstands were sufficient to downgrade circulation patterns to a degree that stimulated widespread evaporite deposition in Laurentia, Baltica, Siberia, and the Australian sector of Gondwana. Some bathymetric changes clearly were local in effect, as brought about by submarine volcanism or by the tectonic elevation or subsidence of the seafloor. Those fluctuations recorded on different paleocontinents during the same interval of geologic time may have been coordinated by minor changes in the size of the surviving Gondwanan ice cap. South American tillites interpreted as Wenlock in age (see above) lend support to this model.

Several small extinction and radiation events in the evolution of nektonic and pelagic organisms appear to be linked to Silurian fluctuations in sea level. Five graptolite radiations are recorded in the Silurian System. Four of these sudden increases in diversity occurred, respectively, in early Aeronian, early Telychian, early Sheinwoodian, and early Gorstian times during or immediately after highstands in sea level. The basal Wenlock (Sheinwoodian) radiation, for example, involves distinct new genera such as *Cyrtograptus* (Figure 21) and as many as 20 new species. Among conodonts, a significant radiation is indicated by species within the *Pterospathodus amorphognathoides* biozone, which straddles the Llandovery–Wenlock boundary and includes the late Telychian highstand. Extinction of key species followed by the origination of several new species during early Sheinwoodian time was one of the most drastic changes in the Silurian conodont succession.

Acritarchs are microfossils that may represent the pelagically dispersed spore cases of benthic algae. Four major turnovers in Silurian acritarch species are recognized. Among those coinciding with highstands in sea level, the mid-Aeronian and early Gorstian turnovers are the most extensive. The various nektonic and pelagic organisms may have been affected by changes in water temperature related to minor episodes of glaciation.

Paleoclimate. If paleocontinental orientations are interpreted correctly and if the assumption is made that atmospheric circulation functioned according to the same basic principles in Silurian time as today, then it is possible to infer general Silurian climatic conditions. A zonally uniform climate is expected in the Northern Hemisphere during the Silurian due to the fact that it was dominated by a north polar ocean. Wind patterns must have included strong polar easterlies at high latitudes, prevailing westerlies at mid-latitudes, and northeast trade winds in the tropics. With the supercontinent of Gondwana centred over the South Pole, climate in the Southern Hemisphere must have been dominated by the interaction of cellular air masses over land and water. The resulting circulation pattern probably was more complex, involving seasonal monsoons.

The Silurian ITCZ

Atmospheric circulation patterns interpreted for an early Silurian summer in the Northern Hemisphere indicate high pressure over the polar ocean with a zone of low pressure around 60° N latitude. Distinct high-pressure cells formed above subtropical oceans, much like the persistent Bermuda high-pressure centre over the present subtropical North Atlantic. Another zone of low pressure formed above the thermal equator, or the region of most intense solar warming. This somewhat migratory zone was the Silurian intertropical convergence zone (ITCZ), where Northern and Southern hemispheric trade winds converged and rising tropical air produced regular cloud cover and precipitation. Mostly, the ITCZ remained near the equator, but it may have migrated slightly to the north in response to strong summer heating on Laurentia, Baltica, and possibly Kazakhstania. This tendency would have been strongest along the eastern margins of tropical continents, where anticyclonic circulation around subtropical highs pulled warm, moisture-laden air northwestward from equatorial oceans. Subtropical high pressure probably spread onto Gondwana, particularly the Australian and Antarctic sectors. A pressure ridge may have merged with these subtropical highs to form a massive cold cell penetrating to higher latitudes over the continental interior of Gondwana. Low-pressure systems over Gondwana's mid-latitude shelf were not unlike the Icelandic and Aleutian lows of today.

Pressure systems should have moved somewhat southward during the Northern Hemispheric winter, particularly the ITCZ, which generally tracks the thermal equator to the Southern Hemisphere during this season. A low-pressure system between Laurentia and Baltica is consistent with the erosion of thick clastics derived from the Taconic and Caledonian highlands. The most significant seasonal variation surely occurred in the eastern Australian and Antarctic sectors of Gondwana, where summer heating abolished winter high-pressure cells and pulled the ITCZ more poleward. This is comparable to today's monsoons, which pull the ITCZ in the opposite direction over the subcontinent of India during the Northern Hemispheric summer. Subtropical highs intensified over the ocean waters of the Southern Hemisphere and probably insulated the arid climate of Western Australia.

Silurian life. Marine benthic invertebrates of the Silurian Period belonged to persistent assemblages or communities that commonly conformed to ecological zonation. Ofle way in which zonation expresses itself is through bathymetric gradients. Paleoecologists studying in Wales, Norway, Estonia, Siberia, South China, and North America have used very similar models to explain the geographic distribution of Silurian communities. Some of these communities were adapted to life under conditions of stronger sunlight and more vigorous wave energy in shallow nearshore waters; others were restricted to darker, quieter environments in deeper offshore waters.

Pentamerid communities

The *Pentamerus* Community (Figure 23) was an early Silurian community dominated by the large-shelled brachiopod *Pentamerus oblongus.* The community often included from 5 to 20 associated species, although enormous monospecific populations sometimes are found preserved in growth position. The *Pentamerus* Community and its slightly older or younger equivalents dominated by similar pentamerid species in the genera *Virgiana, Borealis, Pentameroides,* and *Kirkidium* all occupied a bathymetric zone of medium water depth. These pentamerid communities are known to have lived in sunlit waters because they are associated with robust, calcareous green algae. The waters were not too shallow, however, because pentamerid brachiopods lost their pedicle (the fleshy appendage that tethers the shell to the seafloor) as they matured, and thus unsecured populations were vulnerable to disruption by steady wave activity. The pentamerid communities thrived within a depth range of perhaps 30 to 60 metres. This was below the level of normal (fair weather) wave activity but still in reach of storm waves. At their lower depth limit, the pentamerid communities were out of reach of all but the most intense and infrequent storms.

From E Winson in W S McKerrow (ed), *The Ecology of Fossils*, Gerald Duckworth & Company Ltd

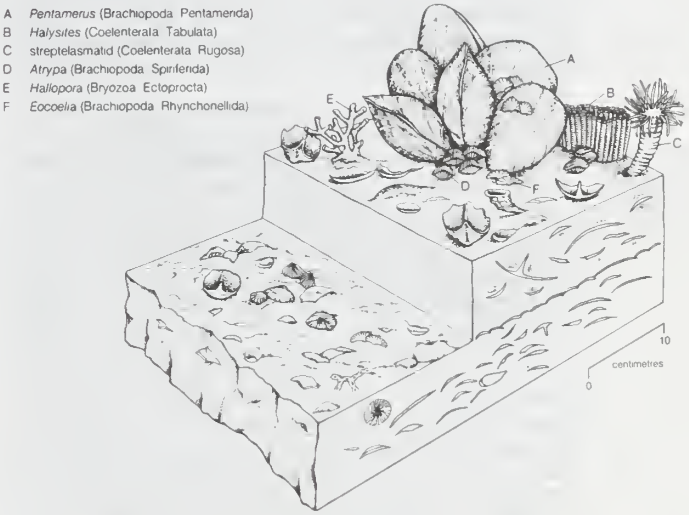

A *Pentamerus* (Brachiopoda Pentamerida)
B *Halysites* (Coelenterala Tabulata)
C streptelasmatid (Coelenterata Rugosa)
D *Atrypa* (Brachiopoda Spiriferida)
E *Hallopora* (Bryozoa Ectoprocta)
F *Eocoelia* (Brachiopoda Rhynchonellida)
A
B
C
D
E
F
10
centimetres
0

Figure 23: An early Silurian *Pentamerus* community.

In regions such as Wales that are characterized by clastic deposition, an onshore–offshore array of five brachiopod-dominated communities may be mapped in belts running parallel to the ancient shoreline. Listed in order from shallowest to deepest position, they are the *Lingula, Eocoelia, Pentamerus, Stricklandia,* and *Clorinda* communities. Below a relatively steep gradient, the centre of the Welsh basin was filled by graptolitic shales. Other areas, such as the Laurentian and Siberian platforms characterized by carbonate deposition, typically developed a continuum of stromatolite, coral-stromatoporoid, *Pentamerus,* and *Stricklandia* communities. *Clorinda* communities were rare in this setting. *Stricklandia* communities sometimes included smaller, less robust individuals of calcareous green algae, indicating a slightly deeper-water environment than that occupied by the *Pentamerus* Community. Coral-stromatoporoid communities (Figure 24), which sometimes formed reef mounds, preferred wave-agitated waters shallower than 30 metres. Much like the reef communities of today, they could not tolerate the more excessive rates of sedimentation typical of clastic settings. Bathymetric

From E Winson in W S McKerrow (ed), *The Ecology of Fossils*, Gerald Duckworth & Company Ltd

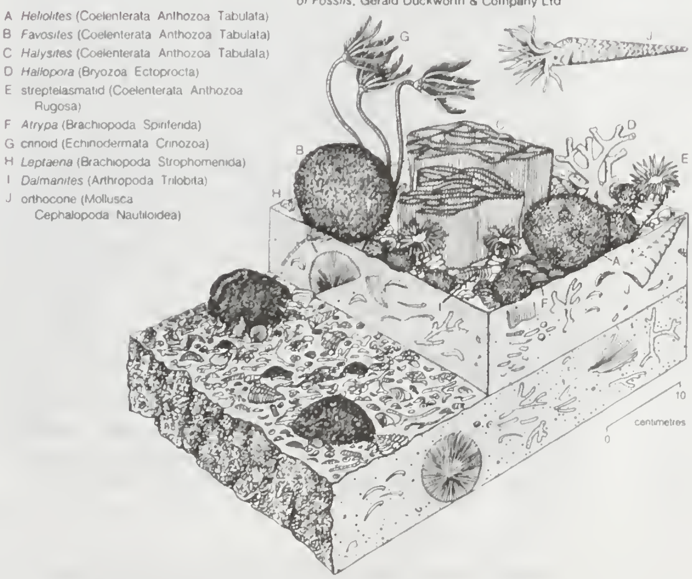

A *Heliolites* (Coelenterata Anthozoa Tabulata)
B *Favosites* (Coelenterata Anthozoa Tabulata)
C *Halysites* (Coelenterata Anthozoa Tabulala)
D *Hallopora* (Bryozoa Ectoprocta)
E streptelasmatid (Coelenterata Anthozoa Rugosa)
F *Atrypa* (Brachiopoda Spiriferida)
G crinoid (Echinodermata Crinozoa)
H *Leptaena* (Brachiopoda Strophomenida)
I *Dalmanites* (Arthropoda Trilobita)
J orthocone (Mollusca Cephalopoda Nautiloidea)
G
J
B
C
D
E
H
A
F
I
10
centimetres
0

Figure 24: An early Silurian coral-stromatoporoid community.

relief on carbonate platforms was very gentle; the full spectrum of available communities usually was expressed over a gradient hundreds of kilometres long. In contrast, the bathymetric gradient on the Welsh shelf was no more than a few tens of kilometres long. Like the *Pentamerus* Community, the other early Silurian communities have ecological equivalents that took their place in later Silurian time. Sea-level fluctuations (see above) are reconstructed by studying community replacement patterns through well-exposed stratigraphic sequences and then comparing the timing of trends on an interregional to intercontinental basis.

Reef mounds (bioherms) provided the Silurian seafloor with an organically constructed microtopography featuring zonations of segregated brachiopods, gastropods, crinoids, and trilobites. The Thornton Reef Complex outside Chicago is an example of a well-zoned Wenlock complex more than one kilometre in diameter. Others are well known from the Silurian of Manitoulin Island (Ontario, Can.), northern Greenland, Shropshire (Eng.), Gotland (Swed.), Estonia, the central and southern Urals of Russia, and Siberia. The most spectacular complex is a 350-kilometre-long barrier reef of late Llandovery–early Wenlock age in northern Greenland. Reefs of all Silurian ages are known, but their development probably reached a climax during Wenlock time. Several thousand bioherms have been recognized from outcrop and subsurface evidence across a tract of 800,000 square kilometres surrounding the Great Lakes region of North America.

Silurian biozonation also is evident on a global scale. The rich benthic faunas just described were tropical to subtropical in distribution. A southern temperate zone, sometimes called the Malvinokaffric Realm, is represented by the low diversity *Clarkeia* (brachiopod) fauna from Gondwanan Africa and South America. A northern temperate zone is represented by the low-diversity *Tuvaella* (brachiopod) fauna mostly restricted to Mongolia and adjacent parts of Siberia. The *Tuvaella* fauna also has been discovered in northwestern China, which apparently represents a more southern extension.

Fish representative of all Silurian ages were widely distributed in marine environments (carbonate and clastic) in a broad belt within the latitudes 40° N and S of the paleoequator. They are known from individual scales as well as from rare body molds. A wide variety of Agnatha (jawless) fish are represented by species belonging to the orders Thelodonti, Heterostraci, Osteostraci, and Anaspida. Fishes with a primitive jaw apparatus are represented by members of the subclasses Acanthodii, Elasmobranchii, and Actinopterygii. Different endemic groups developed in Laurentia (known widely from sites in the Canadian Arctic, the Yukon, Pennsylvania, New York, and especially Scotland), Baltica (especially Norway and Estonia), and Siberia (including adjacent Mongolia).

Vascular land plants

Land colonization by vascular plants was under way during most of the Silurian Period, although activity clearly was restricted to coastal lowlands. Plant megafossils preserved as coalified impressions are fragmentary. Their known distribution includes most of the Silurian continents with limited representation on Laurentia (New York and northern Greenland), Baltica (Avalonian Wales and England, as well as Podolia), the Siberian corner of Sinkiang (northwestern China), and some Australian and North African sectors of Gondwana (Victoria and Libya, respectively). Latitudinal distribution apparently ranged from about 45° N (Siberia) to 30° S (Libya). Species belonging to the genus *Cooksonia* were among the first and most successful vascular land plants found in all the above-cited areas except for northern Greenland and Australia. These plants were small (about six centimetres in height) with smooth, simply branched stems bearing spore sacs at their tips. Photosynthesis took place entirely within the leafless stems. A distinctly endemic group is represented by the genus *Baragwanathia* (Ludlow age) in Victoria, Australia. (M.E.Jo.)

DEVONIAN PERIOD

General considerations. The Devonian Period is the interval of the Paleozoic Era that follows the Silurian and precedes the Carboniferous. It is thought to have covered the span of time between about 408 and 360 million years ago (see Table 4). The rocks that were formed or deposited during this interval make up the Devonian System. Rocks of this system are known on all continents both at outcrop and in the subsurface. Extensive areas of North America, South America, Europe, and the Soviet Union are underlain by Devonian rocks at depth. Subsequent folding has made such rocks common in many of the ancient fold belts.

Paleocontinents of the Devonian

Rocks of Devonian age were deposited after the fusion of the so-called Iapetus Suture had joined the paleocontinents of Laurentia and Baltica (see above) during the Caledonian orogeny. The combined landmass, known as Laurussia or Euramerica, gave rise to widespread areas of continental desert, playa, and alluvial plain depositions. This Old Red Sandstone continental area of what is now Spitsbergen, Greenland, some of Canada's Arctic Islands, Wales and Scotland, and the Baltic Shield forms one of the earliest documented large areas of nonmarine sedimentation. The Old Red Sandstone sectors of eastern and western North America, central and southern Europe, and parts of European Russia are fringed by marine deposits, however.

The present-day southern continents of South America, Africa, India, Australia, and Antarctica were joined together as the enormous continental mass called Gondwana during the Devonian. At this time Gondwana began impinging upon Laurussia. Also, large areas of Asia east of the Ural Mountains were divided into separate landmasses at this point in Earth history. Their distribution is poorly understood, but many of them may have been attached to the margins of Gondwana.

The name Devonian is derived from the county of Devon, Eng. Sedgwick and Murchison proposed the designation in 1839 for the marine rocks they encountered in southwestern England, following the recognition by another British geologist, William Lonsdale, that fossil corals from Torquay in Devon seemed intermediate in type between those of the Silurian System below and those of the Lower Carboniferous System above. This led to the conclusion that the fossil corals were marine equivalents of the terrestrial Old Red Sandstone rocks already known in Wales and Scotland. The recognition that such major paleogeographic differences existed was a great scientific advance, and it was soon confirmed when Sedgwick and Murchison visited Germany and again when Murchison discovered an intercalation of Devonian marine fossils and Old Red Sandstone fish near St. Petersburg in northwestern Russia. By 1843 James Hall of the United States was able to describe equivalent rocks in eastern North America, but precise correlation with European rocks was not achieved until some years later.

Boundaries and subdivisions. During the last half of the 20th century, the International Union of Geological Sciences (IUGS) defined the base of the Devonian System—*i.e.,* the Silurian–Devonian boundary—in a section of an outcrop of sedimentary rock near Klonk, Czech Republic (see above *Silurian Period: General considerations*). A point at La Serre in southern France has been identified as the Devonian–Carboniferous boundary. The base and top of the Devonian System mark the beginning and end of Devonian time.

Devonian series and stages

The characteristics of local sequences have given rise to many different names for subdivisions of Devonian rocks in various parts of the world. For international standardization, however, the IUGS now accepts division of the Devonian System into three series, which are in turn subdivided into stages (see Table 10).

The stage divisions given in the table were agreed upon only recently, so that different terminology may be found in older literature. Each of the stages listed either has been or soon will be defined by a type section or by what is termed a global stratotype section and point (GSSP).

Historically the early Devonian had been divided into the Gedinnian and Siegenian, which were stages based on clastic sections in Belgium and Germany. These sections were difficult to correlate globally, and highly fossiliferous calcareous rocks of similar age in the Czech Republic proved much more applicable. Thus the Czech terms have

Table 10: Subdivisions of the Devonian System

series	stages
Upper Devonian	Famennian (name derived from Famennes, Belg.) Frasnian (from Frasnes, Belg.)
Middle Devonian	Givetian (from Givet, Fr.) Eifelian (from Eifel Hills, Ger.)
Lower Devonian	Emsian (from Ems, Ger.) Pragian (from Prague, Czech Rep.) Lochkovian (from Lochkov, Czech Rep.)

been adopted for international definition. The designation Coblenzian was formerly employed and encompassed both the Siegenian and Emsian stages. Among French-speaking geologists the stage name Couvinian (from Couvin in Belgium) was used in a sense similar to the Eifelian. The new boundary between the Middle and Upper Devonian is somewhat higher than that formerly used in Germany and Belgium. The new definitions reflect the need to establish internationally useful marker levels.

Difficulty in correlating the Silurian–Devonian boundary

Stratigraphic boundaries within the Devonian System are correlated using various fossil groups. In Devonian marine deposits, the conodonts, ammonoids, spores, brachiopods, and corals are particularly useful. In nonmarine deposits such sea-living forms are not found, and freshwater fish and plant spores are employed for correlation. In the past, considerable difficulty was encountered in correlating the Silurian–Devonian boundary, and serious errors were made until recently. This situation resulted because of the misconception that graptolites became extinct at the boundary. It is now known that these invertebrates range almost to the Pragian–Emsian boundary. In areas where graptolites range so high, especially in mainland Europe and what was once the Soviet Union, much miscorrelation occurred. Today the base of the graptolite zone of *Monograptus uniformis* is regarded as the Silurian–Devonian boundary.

The Devonian–Carboniferous boundary similarly had been variously defined in different parts of the world. Although by 1935 it had been agreed to establish the boundary at the entry of the ammonoid *Gattendorfia* using a section in Hönnetal, Ger., French- and Russian-speaking geologists did not universally follow this approach, as they preferred an earlier level. It was subsequently discovered that a gap in the spore record occurs in Hönnetal below the entry of *Gattendorfia*, and the matter has been thoroughly investigated, resulting in a new IUGS definition based on the first appearance of the conodont *Siphonodella sulcata*.

Radiometric evidence has provided various estimates of the time interval represented by the Devonian System. Recent estimates for the base have ranged between 400 and 416 million years ago, and estimates for the top have ranged between 356 and 367 million years ago. (The different estimates carry method errors in excess of these figures.) In practice, it is the evolution of fossil groups that is used to subdivide and correlate rocks of the Devonian System, and at present the conodont and ammonoid scales give about 55 zonal divisions of the system, yielding a relative resolution of about 1 million years in most parts of the system, although in some parts it is considerably better.

Old Red Sandstone continent

Distinctive features. As previously noted, the Devonian System is well represented on all continents. The development of the Old Red Sandstone continent in the area of North America, Greenland, Scandinavia, and the northern British Isles, which were united during Devonian time, gives the first extensive evidence for nonmarine conditions. Because of this, the Devonian is remarkable for its evidence of the colonization of land as well as freshwater rivers and lakes by plants and fish. Both groups existed prior to this time, but they had their earliest extensive evolutionary radiation during the Devonian.

Economic significance. Devonian rocks are locally of economic importance. Marbles of Devonian age have been quarried in France and Belgium. German medieval castles are mostly clad with Devonian slates. In many countries Devonian rocks have provided building stone, refractory and building brick, glass sands, and abrasives. In areas of European Russia and in Saskatchewan, Can., evaporites, including anhydrite and halite, are commercially exploited. Lodes of tin, zinc, and copper occur in several areas where Devonian rocks have been subject to orogenic processes, as in Devon and Cornwall in England and in central Europe. Since the 19th century, oil and natural gas have been produced from Devonian rocks in New York and Pennsylvania. In the 1930s oil was found in Devonian sandstones in the Ural–Volga region and later in the Pechora area of northern European Russia. In 1947 oil was discovered in an Upper Devonian reef at Leduc, Alta.; this was followed by vigorous exploration, and oil production from the area remains significant today.

Devonian rocks. *Occurrence and distribution.* It generally is believed that Europe and North America were united approximately along the present continental slope margins during the Devonian Period. At the close of the Silurian and continuing in the Early Devonian, considerable igneous activity occurred in the belt including New England, Nova Scotia, Newfoundland, Scotland, Scandinavia, and eastern Greenland. With North America and

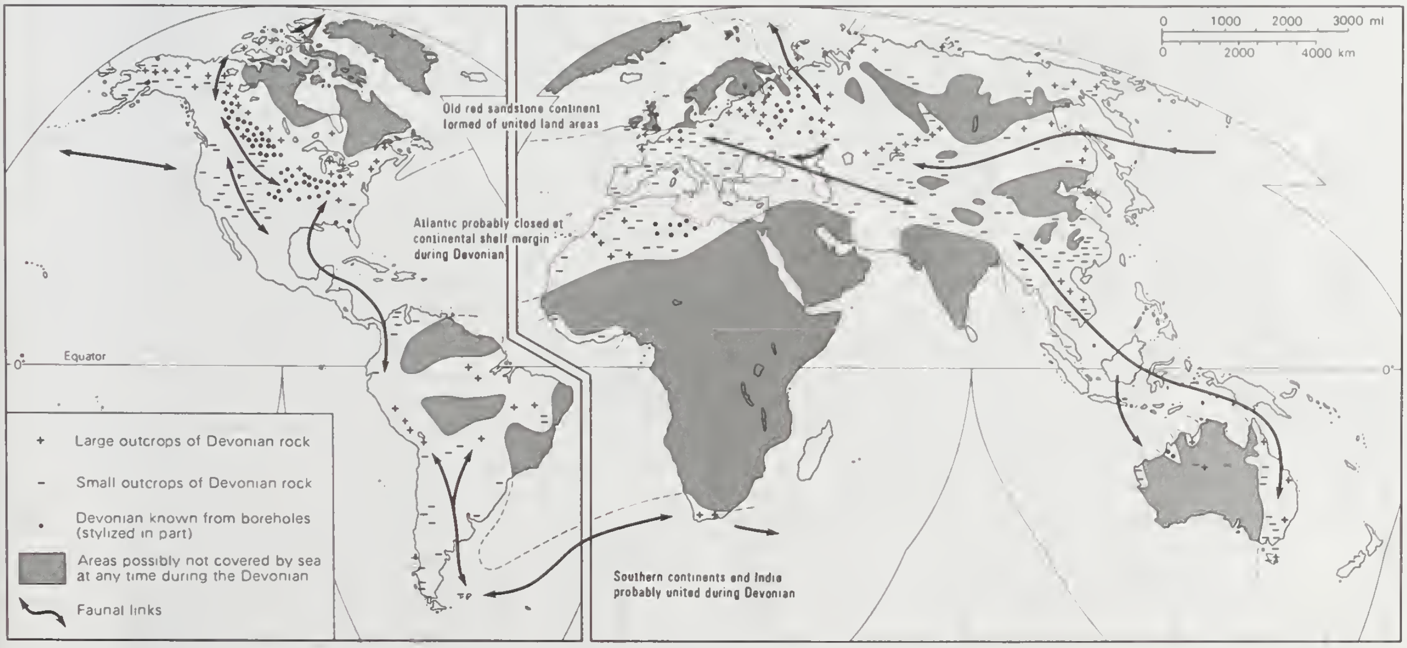

Old red sandstone continent formed of united land areas
Atlantic probably closed at continental shelf margin during Devonian
Southern continents and India probably united during Devonian
Equator
0
0°
0 1000 2000 3000 mi
0 2000 4000 km
+ Large outcrops of Devonian rock
- Small outcrops of Devonian rock
• Devonian known from boreholes (stylized in part)
Areas possibly not covered by sea at any time during the Devonian
Faunal links

Figure 25: Devonian outcrops, inferred land areas, and faunal links.

Table 11: Devonian Stages of the World

	standard stages for the world	Spitsbergen	Scotland Caithness	North Devon	South Devon	Ardennes, Belgium	Rhenish Massif	Czechoslovakia	Russian Platform and Podolia	Southern Siberia Salair	Central and South China	New South Wales, Australia	Western Australia	Bolivia, South America	North American Series
Upper	Famennian		? Upper Old Red Sandstone	Pilton Beds (Pars) Baggy Beds Upcott Beds Pickwell Down Sandstone	slates with volcanics and ostracods	Psammites du Condroz Schistes de la Famennes	Wocklumer Schiefer Dasberg Schiefer Hemberg Schiefer Nehden Schiefer		Dankov Elezk Lebedjan Zadonsk	Podonino Pestchorka	Makunao Tutzutang Changlung-chieh Xikwangshanian	Luton Formation Mandowa Keepit Cong.	Fairfield Group Piker Hills Formation Windjana Reef and Pillara Backreef Nullara Limestone		Chautauquan
	Frasnian			Morte Slates	Saltern Cove Beds Babbacombe Slates	Assise de Matagne Groupe de Frasne	Adorf Kalk		Livny Evlanov Voronezh Rechitsa Semiluki Sargaevo Kyn Paschija	Solomino Hlubokaya Kurlyak Teryochino Vassino	Shetienjiao and Guilin formations Shetienjian	Baldwin ?	Virgin Hills Formation Gogo Formation Pillara Limestone ?		Senecan
Middle	Givetian	Wijde Bay	J. O'Groats	Ilfracombe Beds	Torquay Limestone	Groupe de Givet	Massenkalk Finnentrop Beds Tentaculites Shale	Srbsko Formation	Starooskol Narov and Pjarnu	Izyly Altchedat Sofonino Kerlegesh Akarachin	Dongganling Formation Dongganglingian	Yarrimie	Tandalgoo Sandstone	Collpacnchu Formation	Erian
	Eifelian Couvinian		Thurso Flags Achanarras Passage Beds Wick Flags	Hangman Grits	Calceala Shales	Assise de Couvin	Brandenberg Beds Mühlenberg Sandstone Hobräcke Beds	Choteč Limestone	Lopushansk	Mamontova Shanda	Napiao Formation Yingtangian	Silver Gully		Huamampampa Formation	Ulsterian
Lower	Emsian	Grey Hoek	? Basement Group	Lynton Beds ?	Staddon and Meadfoot Beds	Assise de Bure Hierges Winenne Vireux Emsian	Hohenhöfen Beds U. Ems	Daleje Shales Dalejan Zlickov Limestone Zlickovian	Dniester Series	Belova Salairka Maliy Bachat	Tangding Formation Shipaian	Wogarda Argillite		Icla Formation Upper Santa Rao Formation	
	Pragian	Wood Bay	?			Petigny Saint-Michel Anor Siegenian		Dvorce-Prokop Limestone Koneprusy Limestone Pragian		Krekov	Vilan Formation Yujiangian Nakaoling Formation Nakaolingian		Worral Formation		
	Lochkovian	Red Bay			Dartmouth Slates	Saint-Hubert Oignes Mondrepuits Gedinnian	Bunte Ebb Bredeneck Huinghausen	Kotys Limestone Lochkovian	Ivane Chortkov Borszczov		Lianhuashan Formation Qunzhou Group Lianhua-shanian	Drik-Drik Formation	Carribuddy Group	Lower Santa Rao Formation	
Uppermost Silurian							Köbbinghäuser Schichten	Budnanian	Skalian						

Europe joined as described, the belt thus indicated formed a mountain tract of active uplift. This is the Caledonian mountain belt that resulted from the Caledonian orogeny. The deposits of the Old Red Sandstones appear to be the detritus produced by the erosion of these mountain areas. The marine Devonian rocks of western Canada and those in a belt from Montana to New York in North America, in Europe from Devon to the Holy Cross Mountains of Poland, on the Russian Platform and Novaya Zemlya, and, again, in the Arctic Islands of Canada appear to provide evidence that marine waters encircled the Old Red Sandstone continent.

The accompanying world map (Figure 25) shows the distribution of most of the major outcrops of Devonian rocks. In many areas the Devonian rocks have been much disturbed tectonically by subsequent deformation. These fold belts may be distinguished from cratonic areas where sediments remain much as they were when formed. The main fold belts in North America are the Cordillera (western mountain ranges, including the Rocky Mountains) and the Appalachian belts in the east. In contrast, the Devonian of the Midwest and adjoining areas is flat-lying. In South America, the main fold belt is the Andes and sub-Andes, and east of this line the Devonian rocks are little disturbed. In Australia the main fold belt is in the east from Queensland to Tasmania. In Europe the Armorican fold belt stretches eastward from Cornwall and Brittany. To the south of this line from the Pyrenees to Malaysia, Devonian rocks are caught up in the Alpine-Himalayan fold belt. Similarly, the Devonian of the Ural Mountains is disturbed, whereas to the west, on the Russian Platform, and to the east there is less deformation. In all these cases the folding occurred well after the Devonian, but there is evidence that Devonian sedimentation contributed to the oceanic belts that were sites of the mountain building that occurred later.

In the regions that have suffered severe deformation, the Devonian sediments are frequently metamorphosed into slates and schists and often lose all the characters by which they may be dated. In areas where little change has taken place, all rock lithologies occur, from those characteristic of continental and desert conditions to the varied lithologies associated with shelf and deep-sea accumulation. Contemporary igneous activity is widespread, both in the form of extrusive lavas, submarine pillow lavas, tuffs, agglomerates, and bentonites and also igneous intrusion. Extrusive activity is found in both continental and marine environments, whereas plutonic intrusions are usually linked with areas of uplift such as the Caledonian and Acadian belts of Europe and eastern North America.

For convenience in description this account will commence with a brief review of the European and North African sequences and then pass eastward to Russia, China, and Malaysia. Treatment of the southern continents from New Zealand to South America will follow, and North America will be considered last. Occurrences of various units are presented in Tables 11 and 12.

Europe

A line passing from the Bristol Channel eastward to northern Belgium and Germany roughly demarcates the Devonian marine area south of the Old Red Sandstone continental deposits, which characteristically are red-stained with iron oxide. The continental deposits extend also to Greenland, Spitsbergen, Bear Island, and Norway. The British geologist Robert Jameson coined the term "Old Red Sandstone" in 1808, mistakenly thinking it to be A.G. Werner's "Aelter Rother Sandstein," now known to be of Permian age. The rocks of this wide area have a remarkable affinity in both fauna and rock type and are usually considered to have been united in Devonian times. The relations with the underlying Silurian system are seen in the classic Welsh Borderlands, where the Ludlow Bone Bed was taken as the boundary until international agreement placed it somewhat higher. In Wales, southern Ireland, and the Scottish Lowlands, thicknesses of detritus, chiefly sandstones, accumulated to as much as 6,100 metres in places, and widespread volcanics occur in Scotland. These sediments are rich in fish and plants, as are the eastern Greenland and Norwegian deposits.

Devonian rocks in Devon and Cornwall are mostly marine, but there are intercalations of terrestrial deposits from the north. In northern Devon at least 3,660 metres of shales, thin limestones, sandstones, and conglomerates occur, the latter two lithologies typical of the Hangman Grits and Pickwell Down Sandstones, which are the main terrestrial intercalations. However, in southern Devon reef limestones are in the Middle Devonian, and the Upper Devonian locally shows very thin sequences formed on submarine rises and contemporary pillow lavas in basinal areas. In northern Cornwall both the Middle and Upper Devonian are primarily in slate facies. Fossils found in these rocks have permitted detailed correlations with the Belgian and German sequences.

Devonian rocks of mixed terrestrial and marine type are known from boreholes under London, and these form a link with the Pas de Calais outcrops and to the classic areas of the Ardennes. There, between the Dinant Basin and Namur Basin to the north is evidence of a northward landmass, as in Devon. Both the Lower and Upper Devonian consist of near-shore and terrigenous sediments that reach thicknesses of 2,740 metres and 460 metres, respectively. The Middle Devonian and lower Upper Devonian (*i.e.,* the Eifelian, Givetian, and Frasnian stages, whose type sections are here) consist mainly of limestones and shales and reach at least 1,500 metres in the south. Reefs are especially well developed in the Frasnian and occur as isolated masses, usually less than about 800 metres in length, separated by shales. Equivalents to the north show red and green silts and shales of marginal continental marine type. Because the Belgian Devonian rocks are well exposed along a north–south line, their changes in thickness, lithology, and fauna have been well-documented.

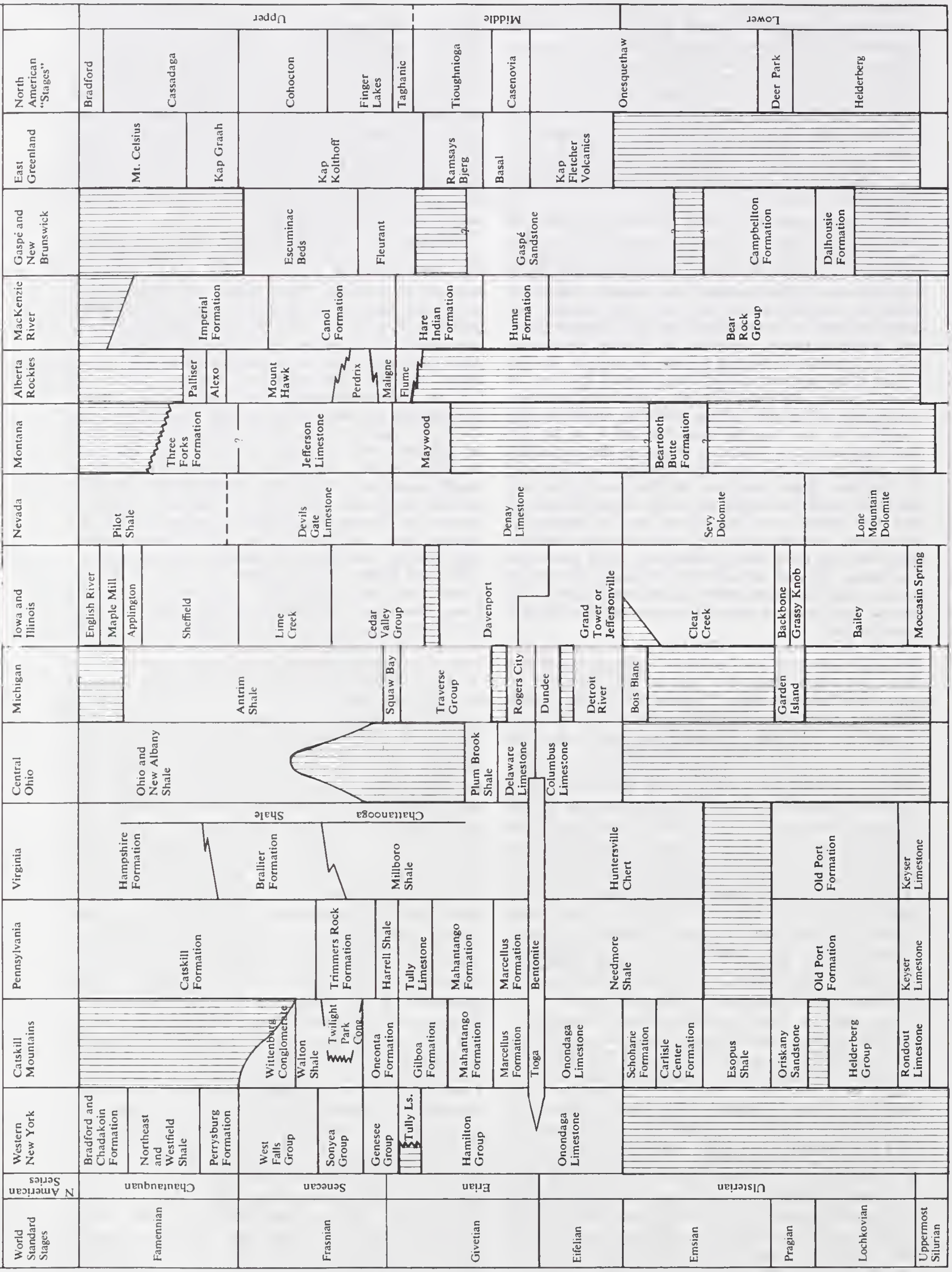
The Eifel forms a natural eastern extension of the Ardennes, and a somewhat similar succession is known. The Lower Devonian is nonmarine, and the Middle Devonian and Frasnian have a poor reef development, but the calcareous shales and limestones carry a rich and famous fauna. The uppermost Devonian is not preserved.

The Rhine valley, along with the Rheinisches Schiefergebirge to the east, has been, since the early days of geology, the subject of extensive study by the numerous German universities that surround it. Again, a northern sediment source generally is indicated, but a borehole near Münster, well to the north, has encountered Middle and lower Upper Devonian marine limestones. To the south also, approaching the Hunsrück-Taunus mountains, there is evidence of a landmass. Between these areas a rich Devonian sequence is exposed in folded terrain. The maximum thickness is 9,140 metres. The Lower Devonian consists of slates and sandstones. The slate has been much worked to clad houses and castles. A ledge of Emsian sandstone in the Rhine gorge is the setting for the Lorelei legend. Limestones are common in the Givetian and are termed Massenkalk. Middle and Upper Devonian areas of thin sedimentation, as in Devon, are interpreted as deposits on submarine ridges. These are commonly nodular limestones rich in cephalopods that occur between thick shale sequences. Evidence of volcanic activity is common, and this has been invoked to explain the concentrations of sedimentary hematite iron ores in the Givetian and Frasnian. The Harz Mountains show a more calcareous Lower Devonian section. Here copper, lead, and zinc are exploited from lodes in the famous Wissenbach Slate.

A calcareous Lower Devonian succession, the Bohemian facies, occurs in the Prague Basin of eastern Europe. A continuous marine succession formed from the Silurian into the Devonian, and the boundary is drawn at the top of the Přídolí Formation with the crinoid genus *Scyphocrinites.* The overlying Lochkovian and Pragian formations include the Koneprusy Limestone with substantial reefs. The Upper Devonian is not preserved. In Moravia complete successions of calcareous and basinal volcanic sediments occur.

Devonian rocks of a type analogous to those of southern England and the Ardennes crop out in Brittany. Farther south outcrops occur in Spain and Portugal. The successions of the Pyrenees, Montagne Noire, and Carnic Alps include deepwater limestones; and marine deposits are known in the Balkan Peninsula, including Macedonia,

Table 12: Devonian Stages of North America



and Romania. The southern Polish outcrops of the Holy Cross Mountains are especially famous and include a lower marine and continental series with a calcareous Middle Devonian and an Upper Devonian of reefs and shales rich in ammonoids and trilobites.

In Podolia, along the Dniester (Dnestr) River, are fine marine sections going well up into the Lower Devonian and overlain by the Dniester Series of the Old Red Sandstone type. During the entire Devonian the Ural Mountains formed a depressional trough linked northward to Novaya Zemlya and southward to the Crimean–Caucasian geosyncline that, with the southern European outcrops already mentioned, formed part of the original Tethyan sediments of the Alpine–Himalayan fold system

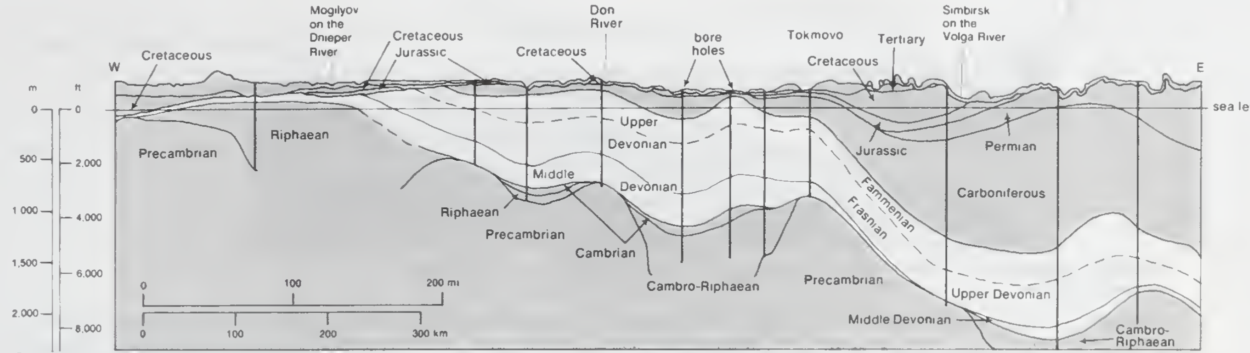


Figure 26: Geologic cross section from near the Polish frontier eastward to the foothills of the Ural Mountains showing subsurface distribution of Devonian rocks.

of the present day. In European Russia, Old Red Sandstone conditions were general, but marine tongues stretched westward from the Urals to reach Moscow in the Middle Devonian and St. Petersburg in the lower Upper Devonian. A remarkable series of boreholes revealed this in great detail (Figure 26), and there is widespread evidence for salt lakes. Apart from the St. Petersburg outcrop and those along the Don River south of Moscow, the salt lakes are known from subsurface data only. Of economic importance here are the Timan-Pechora oil and gas field and the oil and potash of the Pripet Marshes. The North African areas of Algeria and especially Morocco are noted for their wealth of fossils.

Asia

Devonian rocks are widespread in Asia east of the Ural Mountains; however, in Devonian time Asia was composed of separated microcratons or terranes, which appear to have been attached, or adjacent, to the northern margin of Gondwana (see below *Paleogeography*). The coalescence into present-day Asia took place after the Devonian. Devonian rocks are well known fringing the central Siberian craton (a Devonian microcontinent), especially in some of the northern coastal islands, in the Kolyma River basin, and even farther east in Siberia. A particularly good record has been found in Kazakhstan. Devonian rocks occur in the Caucasus and Tien Shan mountains along the southern border of Kyrgyzstan, and there is an excellent carbonate sequence in the Salair and a full marine sequence in the Altai. The Altai–Sayan area contains a wealth of Old Red Sandstone fish and plants.

Scattered Devonian sequences occur in Turkey, Iran, and Afghanistan, but the Himalayan records need revision, as it has now been determined that reported significant fossils are spurious and come from quite different areas. Isolated Devonian rocks are known in Vietnam and Malaysia.

The Greater Khingan Range has a good record of Middle and Upper Devonian marine deposits. China is especially noted for its Devonian rocks; both marine and nonmarine facies occur. Reefs and carbonate deposits also are well developed, and the photographically spectacular sugar-loaf hills near Kuei-lin are of Devonian age. Much research by Chinese geologists since the early 1980s has led to great advances in knowledge of the Devonian in the many outcrops in the People's Republic of China. Devonian rocks in Japan contain the plant genus *Leptophloeum*, which is also widespread in China.

Southern Hemisphere

In New Zealand the Lower Devonian is known in the Reefton and Baton River areas. The brachiopods in the fauna include European elements and have few typical austral types.

Devonian rocks are known in eastern Australia in a belt from Queensland to Tasmania as part of the Tasman geosyncline. Fluviatile sediments are found to the west. Thicknesses of 6,100 metres are known. *Leptophloeum* is found in the Upper Devonian portion. Devonian rocks occur in central Australia in Lake Amadeus and along the western coast in the Carnarvon, Canning, and Bonaparte Gulf basins. Complex facies changes are known, and the Canning Basin reef complexes show every detail of forereef, reef, and backreef structures exposed by modern erosion.

In the Antarctic both marine and continental Devonian occur, the latter rich in fossil fishes of European genera. The marine Lower Devonian shows some affinity with the Bokkeveld in South Africa, which, in turn, has strong links with South America. No Devonian is known in Africa between Bokkeveld and sections in Ghana and northwestern Africa.

Early Devonian marine rocks are well developed in South America, but the Late Devonian is poorly documented. In the western mountains of the Andes and sub-Andes, Devonian remnants are preserved from southern Chile north to Peru, Ecuador, Venezuela, and Colombia. The Devonian rocks of Uruguay, Argentina, and Brazil are thought to represent marine transgression from the west. Both continental and marine faunas have been documented. The fauna of the Falkland Islands as well as of the Paraná and Parnaíba basins includes many genera of brachiopods and trilobites that are common within the circum-Antarctic region but unknown in the Northern Hemisphere. In Venezuela and Colombia, however, faunas of Appalachian type dominate, although austral elements, such as *Australospirifer*, linger.

North America

The Appalachian area of eastern North America shows spectacular and historically famous Devonian rocks first described by James Hall in New York state. A source of sand and other clastics in the east provided a flood of sediment from an eastern land area, which formed the Devonian Catskill Delta that filled a broad sedimentary trough (Figure 27). In the area encompassing Ontario, Michigan, and Indiana, early thin calcareous sequences give way to deeper-water marine black shales, which were formed especially in the area of the Great Lakes and south beyond Indiana. The central area of the United States formed a mid-continental rise during the Devonian, and the Devonian rock record there is thin and incomplete. Devonian rocks are well developed in New Mexico, Utah, Nevada, and north to Montana, where evaporites in the subsurface are known to extend into Saskatchewan. In the mountainous area of the eastern United States, Devonian rocks are scattered and may have coalesced from separate microcratons or microplates over a long period of time. Very thick sequences of Devonian volcanics are known, for example, in the Sierra Nevada of California. In western Canada flat-lying Devonian rocks are well known in the

From M R House, *Continental Drift and the Devonian System*

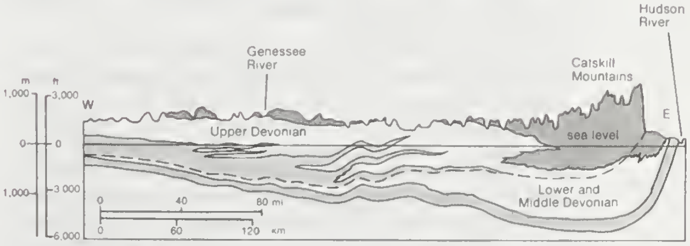


Figure 27: *Cross section along the New York–Pennsylvania border showing changes in Devonian rocks.*
Light gray areas represent red and green shales, sandstones, and conglomerates deposited in fresh or brackish water. Darker gray area represents dark gray or black shales and limestones with marine fossils. White areas indicate gray shales and siltstones with shallow-water marine organisms.

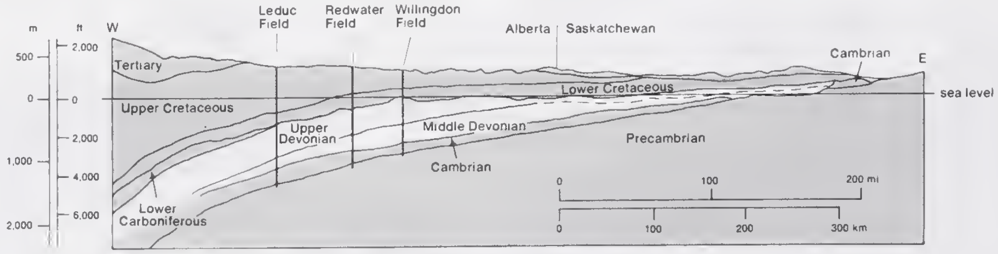


Figure 28: Geologic cross section from the foothills of the Rocky Mountains eastward across the interior plains of Canada showing location of oil fields in the Devonian subsurface rocks.

subsurface of Saskatchewan, and in Alberta they include oil-bearing Devonian reefs (Figure 28). Devonian reef complexes also occur along the Canadian Rocky Mountains. Involved in the thrusting of the Rockies, they can be seen in Alberta's Banff and Jasper national parks. In more scattered outcrops to the east, it would appear that deeper-water facies are represented. Following the discovery of oil in a Devonian reef at Leduc, Alta., much detailed exploration was undertaken. Rocks of Devonian age are widespread from there northward to the Canadian Arctic islands and Alaska. Their faunas show many similarities with those of Europe.

Types. A wide range of terrestrial and marine sediments of Devonian age are known internationally, and there is a corresponding variety of sedimentary rock types. Devonian igneous activity was considerable, albeit localized. The Old Red Sandstone continent is thought to have been near-tropical and sometimes arid. Playa facies, eolian dunes, and fan breccias are known. Fluviatile sediments, deposited by water under flash-flood conditions, have been identified, and these link to alluvial sediments of broad coastal flats. There are lacustrine deposits of freshwater or supersaline type. Similar facies are known in other continental areas of the Devonian. Similarly, nearshore clastic, delta sandstones and prodelta and offshore mud facies compare with those known in other periods.

Special features

Devonian sedimentary rocks include the spectacular carbonate reef deposits of Western Australia, Europe, and western Canada, where the reefs are largely formed of stromatoporoids. These marine invertebrates suddenly vanished almost entirely by the end of the Frasnian Age, after which reefs were formed locally of cyanobacterian stromatolites. Also distinctively Devonian is the development of locally extensive black shale deposits. The Upper Devonian Antrim, New Albany, and Chattanooga shales are of this variety, and in Europe the German Hunsrückschiefer and Wissenbacherschiefer are similar. The latter are frequently characterized by distinctive fossils, but rarely of the benthic variety, indicating that they were formed when seafloor oxygen levels were very low. Distinctive condensed pelagic limestones rich in fossil cephalopods occur locally in Europe and the Urals; these form the facies termed "cephalopodenkalk" or "knollenkalk" in Germany and "griotte" in France. In former times, the latter was worked for marble. Evaporite deposits are widespread, but coals are rare. There is no firm evidence for glacial deposits. Various types of volcanic rocks have been observed in the areas that were converging island-arc regimes. Some volcanic-ash horizons, such as the Tioga Metabentonite of the eastern United States, represent short-term events that are useful for correlation.

Correlation. Most groups of fossil forms contribute to the establishment of a faunal and floral chronology that enables Devonian rocks to be correlated. For the continental deposits, fish and plant spores are most important. The fish give a very precise zonation in parts of the system. The Baltic Frasnian, for example, can be divided into at least five time zones using psammosteids (Agnatha), thus probably equaling the precision possible for the better-known marine Frasnian sequences. Many problems remain, however, in the correlation of the continental and the marine deposits.

The faunal succession in marine strata has been established for many groups, but only those of significance for international correlations are mentioned here. Traditionally the goniatites and clymenids (ammonoid Cephalopoda) form the standard. The succession established first in Germany by the paleontologist Rudolf Wedekind in 1917 has been found to hold for all continents where representatives have been discovered. The index genera are shown in Table 13. All these genus zones (or *Stufen*) are subdivisible into species zones.

Rivaling the ammonoids in most parts of the Devonian and useful for defining the base of the system are the conodonts. The Late Devonian was characterized by a spectacular evolutionary radiation of *Palmatolepis* and its relatives.

The brachiopods, although more restricted, are also important. This is particularly true of the spiriferids of the Early Devonian and of the entry and evolution of the cyrtospiriferid types in the Late Devonian. The rhynchonellids also are of great value in the subdivision of the Late Devonian. Some brachiopods, however, show diverse distribution patterns. *Stringocephalus,* a well-known Middle Devonian guide fossil in the western United States, Canada, Europe, and Asia, is entirely absent from the rich New York succession; yet *Tropidoleptus,* elsewhere confined to the Lower and Middle Devonian, ranges high in the Devonian of New York. Corals also have been used for correlation, but further work suggests they were particularly sensitive to changing local environments and thus are poor time indicators.

Tables 11 and 12 show rock sequences in many parts of the world and how they are thought to be correlated.

Devonian environment. Rock types of the Devonian System indicate that most environments of the present day were represented but that they were very differently distributed. During Devonian times the equator is thought to have passed across Laurussia, and so the Old Red Sandstone continent was essentially tropical or subtropical. Paleomagnetic evidence, however, is not clear, and various positions for the equator have been proposed. Furthermore, paleomagnetic evidence can suggest former magnetic latitudes, but it does not indicate longitudinal position. For Gondwana, the evidence suggests a pole probably in the South African area during the Devonian.

Paleogeography. A description of the physical geography of the Devonian can be attempted using evidence from paleomagnetism, paleoclimate, paleobiogeography, and tectonic reconstruction. Because the paleomagnetic data for the Devonian remains problematic, recent efforts to elucidate paleogeographic position have concentrated on the rock types associated with particular environments and to a lesser extent on faunal distributional data. Such methods use the distribution of evaporites, shelf carbonates, and hermatypic corals, since the present-day aspects of these deposits have specific climatic constraints.

The reconstruction for the later part of the Early Devonian (shown in Figure 29) reflects one interpretation of continental distribution. There is general agreement that the equator crossed the northern part of the Old Red Sandstone continent during the Devonian but that it migrated southward over the course of the period. This is indicated by the reduction in evaporitic environments in western Canada and the onset of humid and moist conditions in the area of New York. Evidence of nonmarine fish and marine invertebrates provides links across the northern area between Europe, Siberia, and the Canadian Arctic islands. Positioning in relation to Gondwana is more difficult. Some interpretations favour a wide ocean separating Gondwana from Laurussia. This arrangement is thought unlikely because of the remarkable occurrences

Table 13: Devonian Genus Zones

system	international		European ammonoid zones		North American ammonoid zones		conodont zones	North American	
	series	stages						stages	series
Devonian	Upper	Famennian	*Wocklumeria*	*Ac. prorsum*	31	? *Epiwocklumeria* species	*Pro. praesulcata*	Bradford	Chautauquan
				Cym. evoluta					
				Wock. sphaeroides					
				Kall. subarmata					
			Clymenia	*Pir. piriformis* *Orn. ornata* *Pro. acuticostata* *Prot. serpentina*	30	*Cymaclymenia* species	*Pa. expansa*	Cassadaga	
							Pa. postera		
			Platyclymenia	*Pl. annulata*	29	*Falc. bowsheri*	*Pa. trachytera*		
				Prob. delphinus		*Pl. americana*			
				Ps. sandbergeri		*Sp. milleri*	*Pa. maginifera*		
			Cheiloceras	*Maen. pompeckji*	27	*Maen. pompeckji*	*Pa. rhomboidea*		
				Cheil. curvispina	26	*Cheil. amlylobum*	*Pa. crepida*		
					25	*Au. clarkei*	*Pa. triangularis*		
		Frasnian	*Manticoceras*	*Cr. holzapfeli*	24	*cf. Cr. holzapfeli*	*Pa. gigas*	Cohokton	Senecan
					23	*Mant. cataphractum*			
				Mant. cordatum	22	*Mant. rhynchostoma*	*An. triangularis*		
					21	*Be. williamsi*			
					20	*Probelaceras strix*	u	Finger Lakes	
					19	*Probelaceras lutheri*			
					18	*Sand. syngonum*	m *Mes. asymmetricus*		
				Koen. lamellosus *Pett. feisti* ? *Pett. errans*	17	*Koen. styliophilum*	l		
	Middle	Givetian	*Pharciceras*		16	*Koenenites* species	lm		
			lunulicosta	*Ep. peracutum*	15	*Pont. perlatum*	*Pa. disparalis*	Taghanic	Erian
				Phar. arenicum *Phar lunulicosta* *Phar. amplexum*	14	*Ep. peracutum*	*Schm. hermanni* / *Po. cristatus*		
					13	*Phar. amplexum*	u		
			Maenioceras	*Maen. terebratum*	12	*Tom. uniangulare*	m *Po. varcus*	Tioughnioga	
					11	*Maenioceras* species			
				Maen. molarium	10	*Sob. virginiana*	l		
					9	*cf. Maen. molarium*	*Po. ensensis*		
					8	*Tom. arkonensis*		Cazenovia	
					7	*Parodiceras* species			
		Eifelian		*Cab. rouvillei*	6	*Ag. vanuxemi*	*To. kockelianus*		
					4.5	*Cab. plebiforme*			
			Anarcestes	*Pin jugleri*			*To. australis*		
							Po. costatus	Onesquethaw	Ulsterian
				F platypleura	3	*F buttsi*	*Po. partitus*		
					2	*Ag. oliveri*			
	Lower	Emsian	*Anetoceras*	*An. lateseptatus*	?1	*An. praecursor*	*Po. patulus*		
				Sell. wenckenbachi			*Po. serotinus*	Deer Park	
				Mim. zorgensis			*Po. inversus*	Helderberg	
				An. hunsrueckianum			*Po. gronbergi*		
				no ammonoids known			*Po. dehiscens*		
		Pragian					*Po. kindlei*		
							Eo. sulcatus		
							Pod. pesavis		
		Lochkovian					*Oz. delta*		
							Oz. eurekaensis		
							Ic. hesperius		

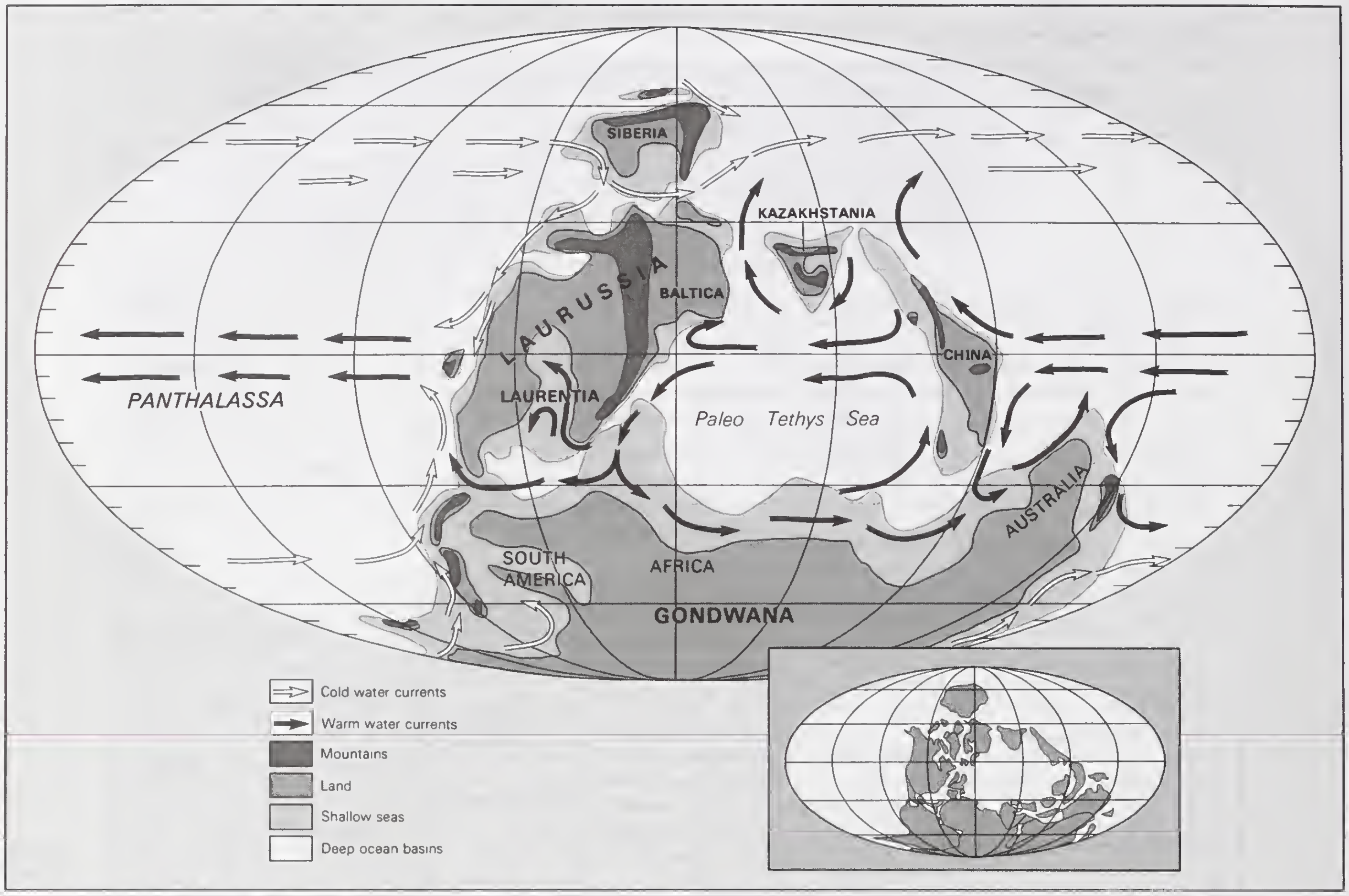


Figure 29: Distribution of landmasses, mountainous regions, shallow seas, and deep ocean basins during Early Devonian time. Included in the paleogeographic reconstruction are cold and warm ocean currents. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset at the lower right.

Adapted from C.R. Scotese, The University of Texas at Arlington

of similar corals, brachiopods, and ammonoids between eastern North America, Morocco, and Spain. Yet, even if they were close together, precise positioning is a matter for dispute, and using the preceding argument some would have North Africa adjacent to the eastern North American seaboard.

There is general agreement that the southern continents of today were united during the Devonian along the lines of their present-day continental slopes. Paleomagnetic evidence is, however, inconsistent on the position of the South Pole; some suggestions favour central South America, while others advance positions in South Africa or sites off the southeast coast. The late Devonian reef developments in Western Australia suggest a near tropical site.

During the Devonian, Asia was composed of many separate microplates that are now joined together. Of these, Siberia and Kazakhstania began fusing during the late Devonian and later joined Laurussia, forming the Ural Mountains along the junction. The positions of the other microcontinents are rather uncertain, but many of them were probably either attached or adjacent to the northern margin of Gondwana and migrated north to fuse with growing Asia at several junctures during the later Phanerozoic.

Tectonic events

Significant geologic events. The union of Laurentia and Europe during the Caledonian orogeny near the beginning of the Devonian Period established a mountain chain that traversed at least from Greenland and western Scandinavia, through Scotland, Ireland, and northern England, including eastern North America especially east of the Hudson River, and continued south to the fringes of western North Africa. Radiometric dating of granitic intrusions yields ages in this belt of between about 430 and 380 million years. The igneous activity that produced such intrusions constituted the final stages of subduction and obduction (*i.e.,* overthrusting of the edge of one lithospheric plate over another at a convergent boundary), leading to the union of the constituent parts of the Old Red Sandstone continent. Considerable extrusive and intrusive volcanic activity would have been associated with the Caledonian orogenic belt. Clastic material from the belt dominated the European Lower Devonian but was local and limited after that point. In eastern North America similar activity near the Silurian–Devonian boundary was followed by renewed activity during the Middle Devonian that was associated with the Acadian orogeny and the commencement of the Catskill Delta. The easterly derived fan clastics of the latter are increasingly dominant eastward across New York state, and its mostly nonmarine alluvial rocks are best seen in the Catskill Mountains near Albany.

It is clear that there was probably easterly directed subduction in western North America during the Devonian. Relics of this process are to be sought in the Cordilleran Mountain chain as discrete terranes that were accreted to the continent during or after the Devonian. The clearest evidence is from the mid-Famennian Antler orogeny, during which a tectonic event that resulted in clastic material spreading eastward is well documented, especially in Nevada.

Sedimentary events

In recent years it has become increasingly apparent that there were specific periods of sedimentary perturbation during the Devonian Period. A dozen or so such "events" have been identified (Table 14). An associated transgression, regression, or transgression/regression "couplet" of brief duration is often involved. Some of these were accompanied by short-term deposition of anoxic (*i.e.,* oxygen-depleted) black shales or limestones. Many are quite widespread internationally, and while eustatic sea-level changes were usually involved, none of the sedimentary indicators appear to be global. Some are associated

Table 14: Periods of Sedimentary Perturbation

	stage	zone	"events"	faunal guides
		SU		*Gattendorfia*
Devonian	Famennian	PR	▮ Hangenberg	*Acutimitoceras*
		EX		
		PO	▮ Annulata	*annulata*
		TR		
		MA	▮ Enkeberg	(clymeniids)
		RH		
		CR	▮ Nehden	*Cheiloceras*
		TR		
	Frasnian	GI	▮ Kellwasser	*Crickites*
		TR		
		AS	▮ Frasnes	*Manticoceras*
	Givetian	DIS		
		H/C		
		V	▮ Taghanic	*Pharciceras pumilio*
		EN		
	Eifelian	KO	▮ Kačák	*rouvillei, otomari*
		AU		
		CO		
		PA	▮ Choteč	*jugleri*
	Emsian	PAT		
		SE		
		LAT	▮ Daleje	*elegans*
		GR		
		DE	▮ Zlichov	*Anetoceras*
	Pragian	KIN		
		SU		
		PE	▮ L. Pragian	
	Lochkovian	DE		
		EU		*uniformis*
		HE	▮	

with the extinction of important groups of fossil organisms. Two are especially important in this regard, the double Kellwasser event and the Hangenberg event.

All the so-called events listed in Table 14 are associated with the loss of certain fossil groups. The pair of Kellwasser events in the late Frasnian show a staged extinction of many groups, especially colonial rugose corals, stromatoporoids, and numerous varieties associated with carbonate environments, including orthid, pentamerid, and atrypid brachiopods, and a large number of trilobite groups. This extinction event has been interpreted as having resulted from one of the following phenomena: a global deepening of the sea by transgression leading to the destruction of reefs and hence of many of their associated fauna; a widespread development of anoxia following transgression and regression; or a meteoroidal or cometary impact and resultant dust clouds. The extinctions of the Hangenberg event at the close of the Devonian show the loss of the clymenid ammonoids and of the phacopid trilobites. Again, there is a link with anoxic conditions, as attested to by the Hangenberg Shale in Europe and by the Exshaw Shale and its equivalents in the western United States and similar lithologies recognized in China.

Paleoclimate. There is evidence for most of the expected climatic belts during the Devonian. Evidence of glacial deposits in the Devonian is questionable, however, and it is clear that, if polar ice caps did exist, they were very much smaller than they are today. It follows that the Earth was warmer during Devonian time than at present.

Prevalence of warm and equable climates

The wide distribution of evaporite basins in the Northern Hemisphere, of coals in Arctic Canada and Spitsbergen, of desert conditions, and of widespread marine faunas and carbonate reefs suggests that warm and equable climates covered large areas. Nevertheless, from New York, where rich forests grew, *Callixylon* trunks with annual rings typical of the seasonal growth of higher latitudes are known. Studies of growth lines on Devonian corals indicate that the number of days in the Devonian year was on the order of 400 and that the lunar cycle was about 30½ days. The wide distribution of salt deposits suggests high evaporation and warm climates in many areas from western Canada to Ukraine and Siberia and, again, locally in Australia.

Devonian life. A highly varied invertebrate fauna derived from that of the Silurian continued in the Devonian, and most ecological niches of shallow and deep marine water were exploited. The remarkable proliferation of primitive fish, which has given the period the name the "Age of Fishes," occurred in both fresh and marine waters. Derivation of carnivorous fish from mud-eating forms occurred early in the period, and the tetrapods were derived from the fish near the close of the period. Also remarkable is the rise to dominance of the vascular plants. By the mid-Devonian the first tree forests are known in place, but rich groves must have occurred earlier to provide the widespread plant debris.

Invertebrates

The Devonian invertebrate faunas are essentially of the type established by the Ordovician. In nearshore sandy and silty environments bivalves, burrowing organisms, brachiopods, and simple corals abounded. In offshore reefs, free from land detritus, biostromes and bioherms flourished, rich in corals, stromatoporoids, crinoids, brachiopods, trilobites, gastropods, and other forms. In deeper waters the cephalopod goniatites, one of the few new groups to appear, were abundant; and there is evidence that the surface levels of these deep waters were occupied by small dacryoconarids of uncertain affinity and by ostracods (arthropods) later in the period.

Both Foraminifera and Radiolaria among the Protozoa are well known, and sponges were locally abundant; the famous dictyospongoids of New York are an example.

The corals and stromatoporoids among the coelenterate Hydrozoa were extremely important in the reef facies. Elsewhere, only simple corals are frequently found. The limestone-reef and forereef facies and biostromal limestones are known in many areas of the world. The corals include tabulate corals, such as *Favosites* and *Alveolites,* but especially rugose corals, which have been used to establish correlations. *Amphipora* is a common rock-building type in the mid-Devonian of the Northern Hemisphere, and its twiglike form produces a "spaghetti" or "vermicelli" rock.

Bryozoa were especially common in shallow shelf seas of the period, and rich faunas are known from North America. Both stony (trepostomatous) and netted forms occurred, but the latter, the fenestellids, became important during the period.

The brachiopods of the Devonian show great diversity. The spire-bearing spiriferoids were perhaps the most common and have been used for zonation. Two groups of importance emerged during the Devonian: the loop-bearing terebratulids and the spiny, mud-dwelling productids. At the same time, a number of groups became extinct, including various orthids and the pentamerids.

Molluscan groups were well represented. The marine clams (Bivalvia) increased greatly during the period, especially in the nearshore environments. The earliest freshwater bivalves appeared in the Late Devonian. The gastropods were well diversified, particularly in calcareous environments, but much less than in later periods. The Scaphopoda first appeared here. A significant Devonian event was the origin of the ammonoids from their continuing nautiloid ancestors. In the chambered shell of the ammonoids, the siphuncle is ventral or outermost in position (except in Late Devonian clymenids), and the septa commence the elaborate folded patterns that culminate in the ammonites of the Mesozoic. From their appearance, probably in the Emsian, the evolution of the goniatites and later ammonites allows a detailed zonal subdivision to be established through the end of the Cretaceous. Devonian goniatites have been found on all continents except Antarctica.

Among the Arthropoda the giant Eurypterida are found in the Old Red Sandstone facies. Some were predacious carnivores and probably lived on fish. The first insect, a supposed collembolan, has been recorded from the Devonian of Russia and other areas of the former Soviet Union. Ostracods were locally very abundant; benthic forms occur in shelf-sea deposits and planktonic forms in the Upper Devonian, where their remains form the widespread ostracod-slate facies or cypridinenschiefer. The trilobites were well developed in terms of size (some up to 61 centimetres long), variety, and distribution. Nearly all have clearly established Silurian ancestors. The most common were the phacopids, which exhibit a curious trend toward blindness in the Late Devonian. Almost all the diverse Lower Paleozoic trilobite stocks that entered the period were extinct before the close, and only the proetaceans survived into the Early Carboniferous.

Among the Echinodermata, holothureans, asteroids, and ophiuroids are known, but they are rare. Crinoids were abundant, including free-living types with grapnel-shaped anchors. The blastoids diversified considerably, but the cystoids did not survive the period.

Conodonts had perhaps their greatest diversification during the Late Devonian and have proved of major importance for correlation.

Vertebrates

Many groups of Devonian fish were heavily armoured, and this has led to their good representation in the fossil record. Fish remains are widespread in the Old Red Sandstone rocks of Europe, especially in the Welsh Borderland and Scottish areas; these are mostly associated with freshwater or estuarine deposits. In other areas marine fish are known, and some of these, such as *Dunkleosteus* (*Dinichthys*) from the Upper Devonian of Ohio, may have reached nine metres in length.

The earliest fish, comprising the Agnatha, were without jaws and presumably were mud-eaters and scavengers. These are usually called ostracoderms. Some, such as the osteostracan cephalaspids, had broad, platelike armour of varied form; and the brain and nerve structures in some of these are well known. The anaspids also were covered with armour in the form of scales. The heterostracans, which include the oldest known fish, have an anterior armour basically of upper (dorsal) and lower (ventral) plates; *Pteraspis* is an example. The Early Devonian saw the entry of jawed forms or gnathostomes, and the armoured forms of these, the Placodermi, characterize the period. The arthrodires with a hinged frontal armour in two portions, and the grotesque antiarchs belong here. The close of the Devonian saw the diminution and extinction of most of these groups, but several other groups continued and have a significant later history. Sharklike fish, the Chondrichthyes, have been found in the Middle Devonian. The bony fish, or Osteichthyes of current classification, include the climatioid acanthodeans, which had appeared before the period began, but the lungfish (Dipnoi), the coelacanths, and the rhipidistians made their first appearance during this time. The last group is thought to have given rise to the four-footed amphibians as well as to all other higher groups of vertebrates.

Plants

In the history of vascular plants, Devonian evidence is of fundamental importance because there was a remarkable initiation of vascular plants of diverse type. Their colonization gave rise to the first forests, such as the rich Gilboa forest of New York, of late Middle Devonian age. Much new information on spores is being provided by palynologists, and this situation may enable the antecedents of the Devonian flora to be established. Evidence of algae is common in the period, Bryophyta are first known here, and Charophyta are locally common. Freshwater algae and fungi are known in the Rhynie Chert of Scotland.

Some supposedly Silurian floras, such as that at Baragwanath, Vic., Australia, are now known to be Early Devonian. The *Cooksonia* Late Silurian record of what is now the Czech Republic seems to be the earliest unquestionable evidence of vascular plants. By the Early Devonian a varied flora was established.

The Psylotophytopsida is the most primitive group of the Pteridophyta; they did not survive the Late Devonian. *Cooksonia, Rhynia,* and others possessing a naked stem with terminal sporangia belong here. In other members, sporangia were borne laterally, but no true leaves were developed, and the branching was often of a primitive dichotomous type. The Psylotophytopsida forms a basic stock from which other groups apparently evolved. *Asteroxylon,* known with *Rhynia* in the Lower Old Red Sandstone Rhynie Chert of Scotland, forms a link with the Lycopsida by having lateral sporangia and a dense leafy stem. This group soon gave rise to treelike forms and later to the important lepidodendrids of the Carboniferous flora. Another apparent derivative, the Sphenopsida, with jointed branches, is represented by *Hyenia* and *Pseudobornia.* The Pteropsida also appeared in the Devonian. Primitive gymnosperms are known and drifted trunks of *Callixylon,* up to 1.8 metres in diameter, occur in Upper Devonian deposits of the eastern United States and the Donets Basin of Russia and Ukraine.

The rich record of land plants may be related to the fact that the Old Red Sandstone represents the first widespread record of continental conditions. However, the primitive nature of the stocks seen and the absence of a long earlier record, even of drifted fragments of vascular plants, suggest that the colonization and exploitation of this environment was a real Devonian event. Fortuitous finds, such as the silicified flora of the Rhynie Chert and the pyritized tissue from the Upper Devonian of New York, have enabled the intimate anatomy of many of these forms to be elucidated in detail equivalent to that of modern forms.

Faunal realms and migrations

There is a marked similarity in the faunas and floras of the continental facies the world over. Recent records from such deposits in China containing Early Devonian genera of the armoured fish *Cephalaspis* and *Pterichthys* or the widespread Australian records of *Bothriolepis,* a Late Devonian antiarch, link closely with the Old Red Sandstone faunas of Europe. Yet when studied in more detail, specific differences become apparent. It has been suggested that the Baltic fish succession is so rich that it must have formed a migration centre. This may be so, but the wide distribution of supposed estuarine and freshwater fish raises many problems. Many of these can be resolved if the continents were closer together during the Devonian than at present.

The marine faunas of the Devonian give little evidence of faunal provinces. It is true that in the Lower Devonian the brachiopod *Australocoelia* has been recognized only in the Antarctic, the Falkland Islands, South America, South Africa, and Tasmania and that *Australospirifer, Scaphiocoelia,* and *Pleurothyrella* share parts of this distribution. These genera are not known in the marine Lower Devonian of northern continents, and this seems to establish an "Austral" fauna of limited circum-Antarctic distribution at this time (if the southern continents were then united as Gondwana). Elements of this fauna are often called "Malvinokaffric" after the Falkland (Malvinas) Islands and the South African Bokkeveld Beds. At other levels in the Devonian, however, provincial distinctions are not apparent, with the exception of local coral provinces that are distinguishable in the areas that once constituted the Soviet Union. (M.R.H.)

CARBONIFEROUS PERIOD

General considerations. The Carboniferous Period is the interval of the Paleozoic Era that succeeds the Devonian and precedes the Permian. In terms of absolute time, the Carboniferous began approximately 360 million years ago and ended 286 million years ago (see Table 4). Its duration of roughly 74 million years makes it the longest period of the Paleozoic and the second longest of the entire Phanerozoic time scale. The rocks that were formed or deposited during the period constitute the Carboniferous System. The name Carboniferous is derived from coal deposits typical of strata in the upper portion of the system throughout the world.

Boundaries and subdivisions. In the early organization and subdivision of Britain's geologic record, coal-bearing strata were combined with a portion of the underlying rock record to be distinguished as the Carboniferous "Order." The Carboniferous initially included the following rock units (listed in ascending order): Old Red Sandstone; Carboniferous, or Mountain, Limestone; Millstone Grit and Shale; and the Coal Measures. The Old Red Sandstone was subsequently placed in the Devonian System, and with time the Upper and Lower subdivisions of the Carboniferous served to differentiate the productive Coal Measures from the barren, predominantly limestone strata below. A similar subdivision was recognizable in North America, and the names Mississippian and Pennsylvanian were proposed as series-rank subdivisions to formalize the same separation of coal-bearing strata from underlying limestones of the Carboniferous System. With the emphasis on the use of unconformities (widespread surfaces of erosion) as a basis for period or system recognition, it was argued that the Mississippian and Pennsylvanian should be separate systems, comparable to the others of

Table 15: Selected Major Divisions of the Carboniferous System and Their Correlation

Permian	Western Europe			former U.S.S.R	China	North America			286
Carboniferous	Silesian — Upper	Stephian	C	Gzhelian	Mapingian	Virgilian	Monogahelan	Pennsylvanian	
			B						
			A	Kasimovian		Missourian			
			CAN.*				Conemaughian		
		Westphalian	D	Moscovian		Desmoinesian			
			C						
			B	Bashkirian	Weiningian	Atokan	Alleghenian		
			A			Morrowan	Pottsvillian		
		Namurian	G_1						
			R_2						
			R_1						
			H_2	Serpukhovian					
			H_1						
			E_2		Tatangian	Chesterian		Mississippian	320
			E_1						
	Dinantian — Lower	Brigantian	Visean	Visean					
		Asbian							
		Holkerian				Meramecan			
		Arundian							
		Chadian				Osagian			
		Courceyan							
			Tournaisian	Tournaisian	Aikuanian				
						Kinderhookian			
Devonian	*Cantabarian								360

(million years ago)

the Paleozoic, because of their unconformable contact. That proposal has been accepted in North America since about 1915, although the U.S. Geological Survey treated the Mississippian and Pennsylvanian as subsystems of the Carboniferous until the 1950s.

The Mississippian and Pennsylvanian have rarely been used outside of North America as systems or lower-ranking subdivisions of the Carboniferous. It has been recognized, at least since 1937, that the Lower–Upper Carboniferous boundary of Europe falls well below the Mississippian–Pennsylvanian boundary in North America. Consequently, neither boundary or subdivision has much stratigraphic significance to the other geographic region. Recently, the Lower–Upper subdivisions of the European Carboniferous were formalized as subsystems with the names Dinantian and Silesian, respectively. In Russia, Ukraine, and other former Soviet republics, a threefold division of the Carboniferous has been employed, with the Lower–Middle boundary approximating the Mississippian–Pennsylvanian boundary and the Middle–Upper boundary falling in the upper portion of the Pennsylvanian.

Carboniferous series and stages

The Carboniferous System has been broken down into series-level subdivisions throughout the world. In Europe the Dinantian, or Lower Carboniferous, is divided into the Tournaisian and succeeding Visean series. The Silesian, or Upper Carboniferous, is divided into the Namurian, Westphalian, and Stephanian series. The British Dinantian has been subdivided further into six stages (in ascending order): the Courceyan, Chadian, Arundian, Holkerian, Asbian, and Brigantian. These stages have for the most part replaced the use of Tournaisian and Visean in Britain, but the Dinantian is still employed as a subsystem. Stages also have been recognized in the Silesian throughout Europe. In the Namurian Series, these stages were originally named for their characteristic ammonoid genus (*e.g., Eumorphoceras* = E stage), with further subdivisions identified by numbered subscripts and letters. This practice has been replaced by the use of geographically named stages, but their letter designations are still commonly given. Currently, the Namurian is divided into the Pendleian (E_1), Arnsbergian (E_2), Chokierian (H_1), Alportian (H_2), Kinderscoutian (R_1), Marsdenian (R_2), and Yeadonian (G_1) stages. Westphalian stages have been lettered A–D. Geographic names adopted for these stages are Langsettian (A), Duckmantian (B), and Bolsovian (C). Westphalian D remains unnamed. The Stephanian is divided into the Cantabrian Stage succeeded by three stages lettered A–C. Barruelian has been adopted recently for Stephanian A. Representative rock exposures, or type sections, have been proposed in Europe for most of these series or stage divisions from which fossils may be collected to fix the time concepts they represent and to provide an objective base for their correlation.

In North America, the Mississippian is divided into the Kinderhookian, Osagian, Meramecan, and Chesterian series. The Pennsylvanian for most of the continent is divided into the Morrowan, Atokan, Desmoinesian, Missourian, and Virgilian series (see Table 15). In contrast, the Pennsylvanian of the Appalachian region, particularly its coal basins, is divided into the Pottsville, Allegheny, Conemaugh, and Monongahela series. In the former Soviet republics the Lower Carboniferous is divided into the Tournaisian, Visean, and Serpukhovian series. The Middle Carboniferous is subdivided into the Bashkirian and Moscovian series, and the Upper Carboniferous is broken down into the Kasimovian and Gzhelian. In China the Lower Carboniferous consists of the Aikuanian and Tatangian series, while the Upper Carboniferous is composed of the Weiningian and Mapingian. Correlation of the boundaries of these various divisions of the Carboniferous is unsettled and is the focus of international research. A suggested correlation chart is given in Table 15.

A proposal for an international classification of the Carboniferous System was presented at the Eighth International Congress on Carboniferous Stratigraphy and Geology held in Moscow in 1975. This proposal, shown in Table 16, represents a significant advance toward standardization of international Carboniferous nomenclature. While final agreement on the names used to designate the various levels of the classification scheme lies well in the future, it has been agreed internationally that the bipartite division of the system would make use of subsystems and that their boundary would fall at a level approximating the appearance of the ammonoid genus *Homoceras,* the conodont species *Declinognathodus noduliferus,* and the foraminifers *Millerella pressa* and *M. marblensis.* A date of 320 million years is generally assigned to this horizon. At the same time, international agreement has been reached on a working definition of the Devonian–Carboniferous boundary. Operationally, this horizon is placed at the base of the appearance of the conodont *Siphonodella sulcata* (in an evolutionary lineage derived from *S. praesulcata*). An international effort is under way to find reference stratotype sections for both the Devonian–Carboniferous and mid-Carboniferous boundaries.

Table 16: International Classification of the Carboniferous System*

system	subsystem	series	stage
Carboniferous	Pennsylvanian	Stephanian	Gzhelian Kasimovian
		unnamed	Moscovian Bashkirian
	Mississippian	Mississippian	Serpukhovian Visean Tournaisian

*Proposed at the Eighth International Congress on Carboniferous Stratigraphy and Geology in Moscow, 1975.
Source: H.R. Lane and W.L. Manger, "The Basis for a Mid-Carboniferous Boundary," *Episodes,* vol. 8 (June 1985).

Varied depositional settings

Distinctive features. The Carboniferous was a time of highly variable depositional settings, exhibiting both shallow-marine and continental environments. In the Northern Hemisphere, the Lower Carboniferous is characterized by shallow-water limestones, while the Upper Carboniferous has cyclic sedimentary deposits, reflecting an alternation of marine and nonmarine conditions and a frequent occurrence of coal swamps. In contrast, the Southern Hemisphere experienced widespread continental glaciations during much of the same interval.

Plants and animals were diverse and had a decisive effect on the accumulation of the Carboniferous sedimentary record. Most Lower Carboniferous limestones are composed of the disarticulated remains of stalked echinoderms known as crinoids. Bryozoans and brachiopods were also common and diverse in the Lower Carboniferous. Coals and associated strata in the Upper Carboniferous contain abundant remains of unusual vascular plants, such as the sphenopsids, lycopods (or lycopsids), and seed ferns. More coal was formed during the Upper Carboniferous than at any other time in the entire geologic record. In the Southern Hemisphere a cold-climate flora, typified by seed ferns, dominated upland environments and became the source of coal deposits as well. Amphibians, which appeared in the Devonian, were joined on land by a great variety of insects. In addition, the first reptiles appeared in the late portion of the Upper Carboniferous.

Pulses of mountain building occurred in the Cordilleran (Rocky Mountain) region of North America, the Hercynides and Ural Mountains of Europe, and in Asia and Africa. The preservation of these physical and biological events is one of the most extensive in the entire Phanerozoic record, and Carboniferous strata are well exposed on all continents.

Economic significance. Much of the recoverable bituminous coal of eastern North America and Europe occurs in large sedimentary basins of the Upper Carboniferous. Smaller coal basins of Carboniferous age are found in North Africa (Algeria and Morocco), northern China, and Korea. Also of considerable economic importance are limestone deposits of Lower Carboniferous age. Such limestones are extensively quarried for building stone, lime, cement, and fertilizer. Other commercially valuable materials derived from Carboniferous rocks include refractory clays and gypsum.

Carboniferous rocks. *Occurrence and distribution—Lower Carboniferous.* The Carboniferous is traditionally broken down into lower and upper divisions that reflect major differences in depositional regimes and biota as well as in age and correlation. The Lower Carboniferous, or Mississippian, is characterized by shallow-water limestones deposited on broad shelves occupying most continental interiors, but particularly in the Northern Hemisphere. By contrast, geosynclinal facies, especially turbidites, formed in deeper troughs along continental margins. Shallow terrigenous clastic facies, such as sandstone and shale, are more poorly developed, and coals are rare.

Types—Lower Carboniferous. Limestones of the Lower Carboniferous are for the most part composed of the disarticulated remains of crinoids. These echinoderms attached themselves to the seafloor by means of a long stalk. The stalk and the flowerlike crown of such animals are composed of plates consisting of single crystals of calcite (calcium carbonate). The crinoids grew in great profusion, forming "meadows" of thousands of individuals. Calcite is extremely stable in warm, shallow marine conditions, and so when individual crinoids died their plates would accumulate on the seafloor as sand-sized sediments, which were subsequently cemented together by calcium carbonate. The crinoid fragments were frequently reworked by currents, and the deposits exhibit both cross-bedding and ripple marks. Deposits of crinoidal limestone approaching 160 metres in thickness are not uncommon for certain intervals of Lower Carboniferous time, particularly in North America. Limestone of this kind is exploited as quarry stone.

Crinoidal limestone deposits

In addition to the crinoidal limestones, oolitic limestones and lime mudstones formed in shallow-water marine environments of the Lower Carboniferous. Ooliths are concentric spheres of calcium carbonate that were inorganically precipitated around a nucleus on warm, marine-shelf margins subject to high wave energy (such as the Bahama Shelf and northern Red Sea today). These deposits also exhibit cross-bedding and ripple marks, which testify to the high-energy conditions. Mixtures of ooliths and abraded fossil fragments, particularly foraminifers, are common in the Lower Carboniferous. Lime mudstones reflect quiet, shallow-water environments, such as are found in Florida Bay and on the west side of Andros Island in the Bahamas, which may have been exposed by tidal change. The carbonate mud is produced through the life cycle of green algae, but fossils are not particularly common in these lithologies. Deposits of these Lower Carboniferous limestones are frequently used as quarry stones as well.

In the upper portion of the Lower Carboniferous, marine cycles are developed, probably reflecting the beginning of mountain building in the Appalachian region of eastern North America. Quartz sandstones typically appeared at the outset of each of these cycles as the seas transgressed across the continental interiors. Shales may succeed the sandstones, followed by limestone development, which suggests the clearing of the water and the establishment of carbonate production by animals and plants.

Limestones of Lower Carboniferous age are typically as-

Chert lenses and beds

sociated with lenses and beds of chert (silicon dioxide). The origin of this chert is somewhat problematic, but it appears to be of either primary or secondary origin. Chert of both origins may occur within a single limestone unit but reflect different times of silicification. Primary cherts form penecontemporaneously with the deposition of the limestones in slightly deeper water settings. Secondary chert forms as a later replacement by groundwater usually involving shallower water deposits. Primary cherts are frequently dark-coloured (flint) and disrupt the bedding rather than follow it. They usually lack fossils. Secondary chert is light-coloured, follows the bedding, and is usually fossiliferous.

In deeper-water carbonate regimes on the margins of the continents, limestones become finer-grained and the biotic component less readily recognizable; crinoids and bryozoans are only a minor component. These deposits are termed the Waulsortian facies, and mounds cored by carbonate mud formed on ramps that extended from the shelf areas into deeper water. Famous exposures of such mounds occur in the Franco-Belgian Basin near Namur in south-central Belgium; southwestern Ireland, particularly County Galway; the Craven basin in northern England; and the Sacramento Mountains of southern New Mexico in the United States. The Waulsortian mounds lack an obvious baffling or framework-building organism that would have formed these cores, although cryptostomous bryozoans have been observed in mounds in New Mexico and crinoid "halos" are associated with mounds in both Europe and North America.

In contrast to the shallow-shelf areas that received carbonate sediments, the deeper intracontinental basins and geosynclines are characterized by terrigenous clastics. In Europe these clastics are termed the Culm facies, and they may also contain volcanic units. In North America the eastern continental margin received thick sequences of coarse clastics derived from the highlands to the east that were formed during the late Devonian. Turbidites of Lower Carboniferous age were deposited in both the Ouachita–Marathon and Cordilleran geosynclines. Turbidites, as well as marine and terrestrial clastics of Lower Carboniferous age, have been reported from South America, Australia, and both North and South Africa. Coal-bearing deposits occur in portions of Scotland, Belgium, China, Russia (Siberia), and Kazakhstan. Evaporite deposits and red beds were formed as part of the Lower Carboniferous record of the Maritime Provinces of Canada. Igneous activity, particularly volcanism and granitoid intrusions, characterizes the Lower Carboniferous record of Australia. The Clyde Plateau lavas of Scotland also are of Lower Carboniferous age.

Type region for the Mississippian subsystem

Correlation—Lower Carboniferous. The type region for the Mississippian subsystem lies in the central Mississippi River valley of the United States. Most of the formations representing the type sequence are found in Missouri, Iowa, and Illinois. The Kinderhookian Series includes the Hannibal Formation and the Chouteau Group. The succeeding Osagian Series includes the Burlington Limestone and overlying Keokuk Limestone. The Meramecan Series includes (in ascending order) the Warsaw Formation, Salem Formation, St. Louis Limestone, Sainte Genevieve Formation, Aux Vases Sandstone, and a portion of the Renault Formation. The Chesterian Series includes (in ascending order) the following formations: upper Renault, Yankeetown, Downeys Bluff, Bethel, Ridenhower, Cypress, Beech Creek, Fraileys, Haney, Hardinsburg, Glen Dean, Tar Springs, Vienna, Waltersburg, Menard, Palestine, Clore, Degonia, Kinkaid, and Grove Church. Other well-known Lower Carboniferous units in North America include the Pocono Group and Mauch Chunk Shale of the Appalachian region; the Fort Payne Chert of Tennessee and Alabama; the Caney and Goddard shales of the Arbuckle region in Oklahoma; the Stanley Shale of the Ouachita Mountains in Arkansas and Oklahoma; the Madison Group and Big Snowy Groups of the northern Rocky Mountains; the Redwall Limestone of the Grand Canyon region; and the Lisburne Group of the Brooks Range in northern Alaska.

Lower Carboniferous units exposed at the famous Avon Gorge section at Bristol, Eng. (in ascending order) include the Shirehampton beds, Lower Limestone Shale, Black Rock Limestone, Gully Oolite, Clifton Down Mudstone, Goblin Combe Oolite, Clifton Down Limestone, Hotwells Limestone, and the Upper Cromhall Sandstone. Other notable Lower Carboniferous formations outside of North America include the limestones at Waulsort and the Black Marble of Dinant in Belgium; the Montagne Noire of the French Massif Central; and limestones in Spain and in the Ural Mountains and the Moscow and Donets basins in Russia and Ukraine.

Occurrence and distribution—Upper Carboniferous. The Upper Carboniferous of the Northern Hemisphere is characterized by deposits reflecting an alternating transgression and regression of the continental interiors by shallow seas. These cyclic sequences of distinct strata known as cyclothems include both terrigenous clastics and limestones. Nonmarine strata typically develop coal beds, and Upper Carboniferous (or Pennsylvanian) cyclothems contain the major portion of world coal reserves. Coal cyclothems are interpreted as having resulted from the widespread continental glaciation reflected in coeval deposits of the Southern Hemisphere. Geosynclinal areas continued to receive clastic facies, particularly turbidites, and brief episodes of mountain building began to markedly affect depositional sequences and their thicknesses.

Coal cyclothems

Types—Upper Carboniferous. Cyclothems occur on a worldwide basis throughout the Upper Carboniferous strata, but they have been most widely studied in North America. The cyclothems display one of two types of development. In the eastern interior of North America, where they were first studied, a single cyclothem might consist of as many as 10 separate beds reflective of a single transgression-regression by shallow seas. The lower portion of the cyclothem is predominantly nonmarine and consists of (in ascending order) sandstone, shale, "freshwater" limestone, underclay (buried soil), and a coal bed. Its upper portion shows evidence of marine conditions and is composed of alternating shale and limestone beds, both of which usually contain fossils. The nonmarine sequence probably represents deltaic conditions associated with the regression that allowed swamp conditions to develop on a delta plain. Transgression began with the shale beds overlying the coal. Rapid regression ends each cycle, which is capped by an unconformity. Most cyclothems are incomplete; they do not exhibit the full sequence of beds.

Cyclothems of the Appalachian Basin coalfields in Ohio, Pennsylvania, and West Virgina typically have a good representation of the nonmarine portion of the sequence with thick coals. These coals formed from the carbonization of plant debris, and it is generally held that 30 centimetres of coal equals the compaction of approximately 1.5 metres of plant material. Some coal beds are remarkably thick. The Mammouth coal bed of the Anthracite Belt (also called the Southern Anthracite field) in eastern Pennsylvania has an average thickness of 11–12 metres throughout its extent. The Pittsburgh seam in western Pennsylvania averages 4 metres thick and is reported workable over 15,500 square kilometres. More than 60 coal seams have been identified in Pennsylvania, although only about 10 have ever been exploited. Coeval cyclothems in the western mid-continent exhibit better development of the marine portion of the sequence and have fewer and thinner coals.

In contrast to the coal cyclothems, predominantly marine intervals of Upper Carboniferous age in the western mid-continent, particularly Kansas, Iowa, and Missouri, feature cylothems with alternating beds of limestone and shale. These cyclothems also reflect transgression and regression by shallow seas, but the lower portion of the cycle is the transgressive event, followed by regression in the upper part. Such a cyclothem begins with a sandy shale containing marine fossils. It is succeeded by dark, carbonate mudstones, which are in turn overlain by black shale. The black shale marks the maximum marine transgression. Above the black shale occur marine carbonate mudstones and grainstones, followed by a return to sandy shale. One striking feature of both coal and marine cyclothems is the tremendous lateral persistency of beds within the sequence. Tracing a single bed from outcrop to outcrop over a distance of hundreds of kilometres would not be uncommon in the mid-continent.

Depositional cycles similar to those of eastern North America can be recognized in Europe, but the distribution of sediments is confined to small isolated basins rather than to a broad cratonic shelf. Nonmarine sequences predominate, and indeed some sequences exhibit no marine influence at all. The positioning of the basins is the result of the folding and faulting associated with the Hercynian orogeny. The Middle and Upper Carboniferous record of eastern Russia is similar to that of North America.

In the Southern Hemisphere, there is a marked cooling event beginning in the Namurian (approximating the mid-Carboniferous boundary), and faunas and floras after that time are highly provincial, impoverished, and adapted to cold climates that persisted into the Permian Period. A single continental mass, Gondwana, existed in the Southern Hemisphere at this time, while several continents, notably Laurussia, Siberia, Kazakhstania, and China (including Southeast Asia), occupied the Northern Hemisphere (see below *Carboniferous environment*). Upper Carboniferous glacial deposits of the Gondwana Realm are characterized by tillites resting on polished and striated bedrock surfaces. Striated cobbles, glaciofluvial deposits, and varved lacustrine (lake) sediments occur over large areas of South America, Africa, India, Australia, and Antarctica. The extensive occurrence of these unusual deposits has been used as an argument for continental drift. Timing of the glacial episodes is still uncertain, and they may have actually begun in the Lower Carboniferous. Furthermore, many glacial advances and retreats occurred, and they were not necessarily simultaneous over the whole of Gondwana.

Gondwanan glacial deposits

Areas marginal to continental masses continued to receive turbidites, particularly in the Ouachita–Marathon region of Arkansas, Oklahoma, and Texas, and the Cordilleran geosyncline in the western United States. Evaporites formed in restricted basins, such as those in Montana–North Dakota (Williston) and the Four Corners area (Paradox), which lay near the Upper Carboniferous equator. Igneous and metamorphic rocks of Upper Carboniferous age reflect the Hercynian orogeny and its equivalents in North America, Europe, and North Africa.

Correlation—Upper Carboniferous: The type region for the Pennsylvanian subsystem is located in central West Virginia. There, the interval is represented by the following groups or formations (in ascending order): Pocahontas, New River, Kanawha, Charleston Sandstone, Conemaugh, Monongahela, and basal Dunkard. Other well-known Upper Carboniferous units in North America include the Jackfork and Johns Valley shales of Oklahoma and Arkansas; the Atoka Formation of Arkansas and Oklahoma; the Supai Group of the Grand Canyon region; the Amsden and Tensleep formations of the northern Rocky Mountains; the Fountain Formation of the central Rocky Mountains; and the Haymond and Gaptank formations of the Marathon region in West Texas. Major coalfields in the United States include the Appalachian region (Pennsylvania, West Virginia, Ohio), Illinois Basin, Mid-continent region (Iowa, Missouri, Kansas), Arkoma Basin (Arkansas and Oklahoma), and north-central Texas.

Type region for the Pennsylvanian subsystem

The type region for the upper part of the Carboniferous in Britain includes the Millstone Grit and the Coal Measures, designations that have been in use since the naming of the system. Local names are applied to specific intervals, and marine horizons, called bands, are named either for their characteristic fossil occurrence (*e.g., Listeri* Marine Band) or for a geographic locality (*e.g.,* Sutton Marine Band). This procedure is followed in most areas outside of North America. Major Upper Carboniferous coalfields occur throughout Europe, especially the central Pennines (Lancashire coal basin), the Scottish border, and southern Wales, in Great Britain; the Franco-Belgian Basin in Belgium; the Saar basin along the border of France and Germany; the French Massif Central (Saint Étienne and Gard coal basins); the Ruhr and Westphalian basins in Germany; the Silesian basin in Poland; and the Moscow and Donets basins in Russia and Ukraine.

Carboniferous environment. *Paleogeography.* Figure 30 shows the assumed positions of the continents during the Upper Carboniferous. As will be noted, the land-sea distribution at the time did not correspond to present-day geography. The bulk of the Earth's landmass was concentrated in the Southern Hemisphere. Relative lithospheric plate movements had brought the continents close together on one side of the globe. Only five major continental masses existed. Laurussia, formed by the joining of Laurentia (principally North America and Greenland) and Baltica (most of northern Europe and Scandinavia) in the Devonian, constituted the single largest landmass in the Northern Hemisphere during the Carboniferous Period. In addition to Laurussia, Siberia, Kazakhstania, and China, which included Southeast Asia, were each separate continents. The enormous continent of Gondwana occupied the Southern Hemisphere. Gondwana combined what are now Africa, India, South America, Australia, the Middle East, and Antarctica. The Tethys seaway separated the landmasses of the Northern and Southern hemispheres near the equator.

Laurussia

During the Lower (early) Carboniferous, Laurussia was apparently more fragmented than Gondwana. Siberia, Kazakhstania, and China occupied positions at high latitudes. Northern Europe and Scandinavia were joined to North America, with what is now northern Canada lying at a mid-latitude position and the United States, Poland, Ukraine, European Russia, Belarus, and adjacent areas near the equator. Gondwana lay entirely in the Southern Hemisphere, although little of the landmass was situated at the South Pole.

By the Upper (late) Carboniferous, plate movements had brought most of Laurussia into contact with Gondwana, thereby closing the Tethys. The two became fused by the Appalachian–Hercynian orogeny, which began in the Upper Carboniferous and continued into the Permian. The United States and northern Europe retained their equatorial position. China and Siberia remained at high latitudes in the Northern Hemisphere.

The distribution of land and sea followed fairly predictable limits. The continental interiors were terrestrial, and no major marine embayments apparently existed. Upland areas of the continental interiors underwent substantial erosion during the Carboniferous. Shallow seas occupied the shelf margins surrounding the continents. It must be remembered that areas that were marginal to the Carboniferous continents may very well have become continental interiors in the present geographic setting (as, for example, much of the United States). Deeper troughs (geosynclines) lay seaward of the continental masses, and their sedimentary record is now represented by mountains.

Significant geologic events. The Carboniferous marks a period of relatively stable crustal conditions between major mountain-building episodes in the Devonian and Permian. Nevertheless, the rocks of both the Lower and Upper Carboniferous show evidence of orogeny and isostatic adjustments in previously formed mountains. Among the major continental masses constituting present world geography, only Antarctica exhibits no recognizable trace of Carboniferous earth movements (tectonics).

Significant mountain building resulted from the collision of Laurussia and Gondwana. The movement of these continents toward each other began in the Lower Carboniferous and continued into the Permian. Their collision resulted in the formation of three mountain belts and ranges essentially at the same time. The Hercynides (occasionally called the Variscan Belt) were formed across southern Europe (including Britain). The Appalachians were formed along the eastern coast of North America, and their counterpart, the Mauritanides, emerged in North Africa. A fourth belt, the Ouachita (Marathon) Mountains, were formed along the southern border of North America. Usually viewed as an extension of the Appalachians, the Ouachita orogeny predates most movements in the Appalachians and is linked to the collision of the southern margin of North America with the South American portion of Gondwana rather than with North Africa.

Formation of the Hercynides, Appalachians, and Mauritanides

Less significant tectonism occurred in the Cordilleran (Rocky Mountain) region during a pulse of the Antler orogeny that elevated the Ancestral Rockies and produced thick clastic wedges, such as the Fountain Arkose. Precordilleran movements occurred in South America, and similar events took place in northern Asia (the Tien Shan, Kunlun, and Timan mountains), Australia, and New Zealand. The Ancestral Urals also were active during the Carboniferous.

Paleoclimate. Carboniferous deposits exhibit steep, latitude-controlled, climatic gradients, reflecting the elongate nature of the continental masses of Laurussia and Gondwana and their positions with respect to the equator, as seen in Figure 30. As would be expected, land areas occupying high latitudes were cold, while those near the equator were warm and moist. Widespread continental glaciation occurred in Gondwana in a manner similar to what took place much later during the Pleistocene Epoch in the Northern Hemisphere. Eustatic rise and fall of sea level resulting from glacial advance and retreat probably caused the cyclothems of the Upper Carboniferous coal swamps that developed in the equatorial region. At their maximum, continental glaciers extended from near the South Pole to nearly 30° S latitude, where subtropical conditions would have more likely prevailed. Coeval continental glaciations did not occur in the high latitudes of the Northern Hemisphere, probably because the landmasses were too small to sustain large ice fields. In addition to the glacial deposits, the floras of the Carboniferous reflect the same climatic gradients. Fossil plants of Carboniferous age found in high latitudes exhibit seasonal growth rings, while those of the presumed equatorial coal swamps lack such rings, just as do modern tropical trees.

Carboniferous life. The Late Devonian experienced major extinctions within some marine invertebrate groups, and Carboniferous faunas reflect a different composition than what had prevailed in the middle Paleozoic. Most notably, reef-forming organisms, such as tabulate corals and stromatoporoids, were limited, and this dearth of framework builders resulted in poorly developed reefs during the Carboniferous. Yet, this period was still one of diverse marine invertebrates. More striking, however, was the remarkable variety of terrestrial animals and plants associated with the coal swamps for which the period is noted. Arguably the most famous Carboniferous fossil occurrence is at Mazon Creek, Ill., in the north-central United States, where ironstone concretions formed around a multitude of different invertebrate and plant specimens. The concretions contain many soft-bodied forms, attesting to the rapid formation and burial of these structures.

Flora and fauna associated with coal swamps

Benthic marine communities in the Carboniferous included a variety of invertebrates. As discussed elsewhere, the crinoids grew in great profusion attached to the seafloor. These animals were solitary suspension-feeders. Their large numbers may have caused baffling of bottom currents, resulting in the deposition of carbonate mud around their bases that developed into mounds on the seafloor. The blastoids, a group of budlike stalked echinoderms related to the crinoids, are abundant in Carboniferous deposits as well. Areas favourable to the crinoids and blastoids were also occupied by the cryptostomous, or lacy, bryozoans. Their fanlike colonies were attached to the seafloor and may have aided in baffling fine sediment. Brachiopods, particularly the winged (or spiriferid) and the spiny (or productid) types, anchored their bivalved shells to the substratum by a fleshy stalk and pumped the water column with a ring of tentacles to recover food in much the same manner as the bryozoans. Brachiopods were particularly common during the Carboniferous, and all orders except the Atrypida and Pentamerida, which became extinct at the end of the Devonian, are found in rocks of the period. Both calcareous and agglutinate foraminifers are represented in Carboniferous deposits, especially limestones. An unusual group of these protozoans, called the fusulinids,

Adapted from C R Scotese, The University of Texas at Arlington

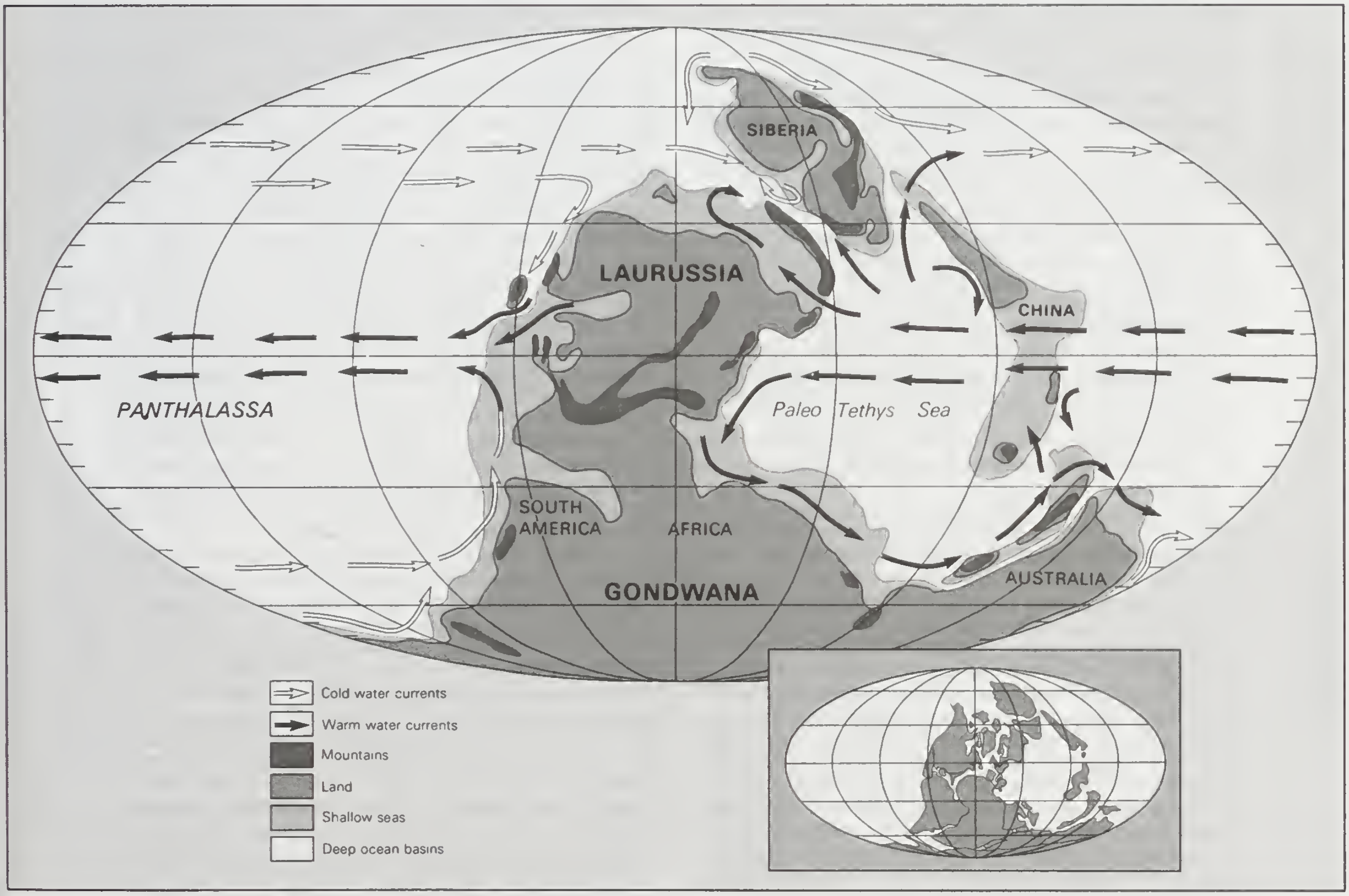


Figure 30: Distribution of landmasses, mountainous regions, shallow seas, and deep ocean basins during the Late Carboniferous. Included in the paleogeographic reconstruction are cold and warm ocean currents. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset.

appear in the rocks of the Upper Carboniferous and dominate the assemblages through the Permian, when they become extinct. The fusulinids secreted a tightly coiled, calcareous test that was chambered. They exhibited rapid evolutionary diversification, and the sequence of various morphologic features reflecting this diversification is used to subdivide and correlate the Upper Carboniferous on an intercontinental basis. In addition to marine invertebrates that inhabited the seafloor, the calcareous algae were well represented. Platelike red varieties were abundant enough to form mounds in the Upper Carboniferous. Benthic organisms on the decline during the Carboniferous were the trilobites, corals, and sponges.

The ammonoid cephalopods, extinct relatives of the chambered *Nautilus,* were common in deep marine waters. They swam by means of jet propulsion and either caught prey or were scavengers. The ammonoids, like the fusulinids, exhibited rapid evolution through the Carboniferous and serve as useful index fossils for correlation of the interval. True nautiloids also were represented in Carboniferous marine environments. While they are not as diverse as the ammonoids, both straight and coiled forms are common as fossils, and some straight forms grew exceedingly large (more than three metres) for invertebrates.

Dominance of vascular plants

Carboniferous terrestrial environments were dominated by plants, ranging from small, shrubby growths to tall trees reaching heights of more than 30 metres. The most important vascular plants were the lycopods, sphenopsids, cordaiteans, seed ferns, and true ferns. Lycopods are only represented in the modern world by the club mosses, but in the Carboniferous they included tall trees that had dense, spirally arranged leaves and spore-bearing organs on their leaves (cones might be present). *Lepidodendron,* with diamond-shaped leaf bases, and *Sigillaria,* with ribs and round leaf bases, were the dominant lycopod genera. They produced fossil logs that measure as much as one metre at their bases. Sphenopsids are trees and shrubs that have a distinctly jointed stem and leaves arranged in spirals from the joints. The horsetail rush (*Equisetum*) is the only living representative, while *Calamites* is the most common Carboniferous genus. Smaller than the lycopods, sphenopsids also occupied somewhat dry, more upland environments. The cordaiteans belong to the gymnosperms and are precursers of the conifers. They also favoured upland environments, grew tall, and had needles and cones like modern conifers, although the group itself has no living representatives. The genus *Walchia* probably formed forested areas much as modern pines do. Seed ferns, or pteridosperms, had fernlike foliage, but reproduced by seeds rather than by spores. They, too, are gymnosperms with no living representatives. The pteridosperms include such trees as *Glossopteris,* a genus that was characteristic of Permian floras from Gondwana; and such low shrubs (usually represented by fragments of their foliage) as *Neuropteris, Pecopteris,* and other forms from Mazon Creek (see above). The seed ferns and true ferns formed the "under" foliage associated with most Carboniferous coal swamps.

The diversity of fish from the Devonian continued into the Carboniferous in both marine and freshwater environments, although the arthodires (armoured, jawed fish) become extinct almost immediately in the Lower Carboniferous. Fully terrestrial environments were populated by a variety of animals. Pulmonate gastropods that respired by lungs arose from aquatic ancestors in the Upper Carboniferous. Insects evolved during the Devonian but diversified through the Carboniferous. No Devonian winged insects are known, and wings must have appeared in the Lower Carboniferous, although no fossil insects are known from that time. Dragonflies and mayflies, which lack the ability to fold their wings, were abundant and attained large sizes by the Upper Carboniferous. More advanced insects that could fold their wings, particularly the cockroaches, also appeared in the Upper Carboniferous, as did the ancestral forms of the grasshoppers and crickets. The earliest land-dwelling scorpions are of Carboniferous age.

While amphibians appeared in the Upper (Late) Devonian, their great radiation took place in the Lower Carboniferous, where they continued as the only terrestrial vertebrates. Although living a semiaquatic life similar to that of modern amphibians, Carboniferous varieties were generally larger by comparison and much more diverse in body plan, even including limbless forms. They may be classified as either labyrinthodonts (named for the infolding of dentine in their teeth), or lepospondyls (small, salamander-like or snakelike forms). The earliest reptiles have been found in Upper Carboniferous sediments filling the external molds of lycopod tree stumps in Nova Scotia. These small (less than 30 centimetres) animals are thought to have become entrapped in the cavities of the rotting trees. They differ only slightly from their amphibian ancestors, principally in the character of their backbone and the number and position of bones in their skulls. Once established, the reptiles diversified rapidly in the Upper Carboniferous, so much so that all the basic skull types except parapsid (*i.e.,* the pattern in which the postorbital and squamosal portions of the skull form a wide cheek) were developed by the end of the period. (W.L.Ma.)

Fossil remains of the earliest reptiles

PERMIAN PERIOD

General considerations. The Permian, the last period of the Paleozoic Era, began about 286 million years ago and ended 245 million years ago, extending from the close of the Carboniferous to the outset of the Triassic (see Table 4). The changes in the general distribution and types of sediments and fossils during these 41 million years suggest that the Permian was a time of progressive climatic shifts that resulted in major environmental challenges to marine and terrestrial life. The rocks that formed during this period make up the Permian System.

The history of the identification and acceptance of the Permian Period by geologists is in many ways the account of good deductive reasoning, a determined scientist, and an opportunity that was exploited to its fullest. R.I. Murchison, also known for his studies of the Devonian and Silurian (and Ordovician) periods (see above), had been aware that the Coal Measures, or Carboniferous System as these rocks became known, in northern England and in Germany were overlain by red beds and dolomitic, poorly fossiliferous limestones that had major rock intervals missing at their base and at their top (Figure 31). Murchison reasoned that somewhere, perhaps outside of northwestern Europe, a more complete stratigraphic succession would fill in these sedimentary gaps and would contain a more complete and better-preserved fossil assemblage. As was discussed earlier in the article, Murchison undertook expeditions in 1840 and 1841 to compile a preliminary synthesis of the geology of European Russia. He traveled extensively throughout the area and met with local geologists and paleontologists to discuss the geologic features of the various districts. Along the western flanks of the Ural Mountains, Murchison recognized that Carboniferous beds were overlain by a well-developed succession of rocks that included rocks equivalent in age to those problematic red beds and dolomitic limestones of northwestern Europe and filled the missing gaps below and above those sediments. He named these rocks the Permian System after the district of Perm, where the succession was particularly well developed.

Early investigations

In his 1845 publications Murchison included the red beds and evaporite beds now referred to as the Kungurian Stage in the lower part of his Permian System and also incorporated the nonmarine beds of the Tatarian Stage in its upper part. The upper portion of these nonmarine beds was subsequently shown to be of Early Triassic age. The Kazanian Stage in the middle is a close lithologic and age equivalent of the Zechstein of northwestern Europe.

Later work by other geologists on the Russian Platform and Ural foothills demonstrated that the clastic beds considered by Murchison to be equivalent to the Lower Carboniferous Millstone Grit were considerably younger and, at least in part, lateral facies of the lower beds of the Kungurian. These and some of the limestone-bearing beds at their base were called the Artinskian Stage. Later, the limestone-bearing lower part was studied in more detail and now has been divided into the Asselian and Sakmarian stages. The Permian succession in its type area as it is presently subdivided is shown in Figure 31.

Table 17: Regional Stages of the Permian System

Adapted from C.A. Ross, *Permian*, in R.A. Robison and C. Teichert (eds.), *Treatise on Invertebrate Paleontology.*

		Pamirs	China		Japan	Australia (Sydney Basin)		North America (West Texas) reference sections		Russian Platform and Southern Ural Mountains type sections			western Europe reference sections		Transcaucasus		Salt Range	
		Lower Triassic	Lower Triassic		Lower Triassic	Lower Triassic		?Middle Triassic ?		Vetuluzhian			Lower Triassic ?		Induan		Mianwali Formation Kathwai Mbr.	
Permian	Upper	Pamirian	Lopingian	Changhsing Formation	Mitaian ?	Newcastle Coal Measures ?		hiatus ?		Tatarian					Dzhulfian	Dorashamian	? Chhidruan	
				Wuchiaping Limestone	Kuman	Tomago Coal Measures		Ochoan					?			"Chhidruan"		Chhidru Formation
		?	?					?		?			?			Araksian		?
		Murgabian	Maokou Limestone (Maokouan)		Akasakan	Maitland Group	?	Guadalupian	Capitan Limestone (Capitanian)	Kazanian			Thuringian	Zechstein	Khachik Formation		Kalabagh Mbr.	
							Muree Formation		Word Formation (Wordian)						Gnishik Formation		Wargal Limestone (Middle Productus Limestone)	
		Kubergandinian										Ufimian		Kupferschiefer				
		?	?							?		?						
	Lower	"Artinskian"	Chihsia Formation (Chihsian)		Nabeyaman	Branxton Formation	"Fenestella Zone"	Leonardian	Road Canyon Formation (Roadian)	Artinskian	?	Kungurian	?		Lower Permian		Amb Formation (Lower Productus Limestone)	
									Cathedral Mountain Formation		Baigendzhinian	Saraninian	Saxonian (Oberrotliegende)					
												Sarginian						
						Greta Coal Measures						?	?					
					Sakamotozawan	Farley Formation			Skinner Ranch Formation		Aktastinian	Irginian ?						
		?								?		Burtsevian	?					
		"Sakmarian"	Maping Limestone (Mapingian)			Rutherford Formation		Wolfcampian	Lenox Hills Formation	Sakmarian	Sterlitamakian		Autunian (Unterrotliegende)		?		Sardi Formation (Lavender Clay)	
						?					Tastubian						Warchha Sandstone (Speckled Sandstone)	
						Allandale Formation ?											Eurydesma-Conularia beds ?	
						Lochinvar Formation			Neal Ranch Formation	Asselian							Tobra Formation (Talchir boulder beds)	
						?											?	
Carboniferous					Hikawan			Virgilian beds		Orenburgian			Stephanian				?	
										Gzhelian								

The Permian was welcomed by many geologists as filling an obvious need for a system of rocks above the Carboniferous and below the Triassic. On the other hand, these latter two systems were well known, long studied, and established in what had become the fertile cradle of geologic thought in northwestern Europe. The type Permian System was distant, sketchily described, and poorly understood. There evolved a strange situation in which textbooks and most geologists widely accepted the Permian System after it was proposed; however, officially the U.S. Geological Survey did not adopt the system as such until 1941, and then only after a symposium organized by the American Association of Petroleum Geologists established North American standard reference sections for the Permian consisting of four series—namely, the Wolfcampian, Leonardian, Guadalupian, and Ochoan—on the basis of the succession in West Texas and New Mexico. The Wolfcampian Series was carried as "Permian?" until 1951 because of continuing controversies in defining the base of the Permian in what was then the Soviet Union.

Boundaries and subdivisions. During the 1960s attempts were made to unify the nomenclature within the Permian System. The system was subdivided into a Lower Permian Series and an Upper Permian Series, and to these were assigned stages with regional names (Table 17). The regional stages are necessary and important because they are based on strongly provincial faunal zonations in stratigraphic successions that differ markedly from one region to the next. This means that within a single region or faunal province the similarity of the succession of fossils and patterns of rock deposition permits ready age correlations. The age correlations from one region to the next are more difficult and open to more questions.

Permian series and stages

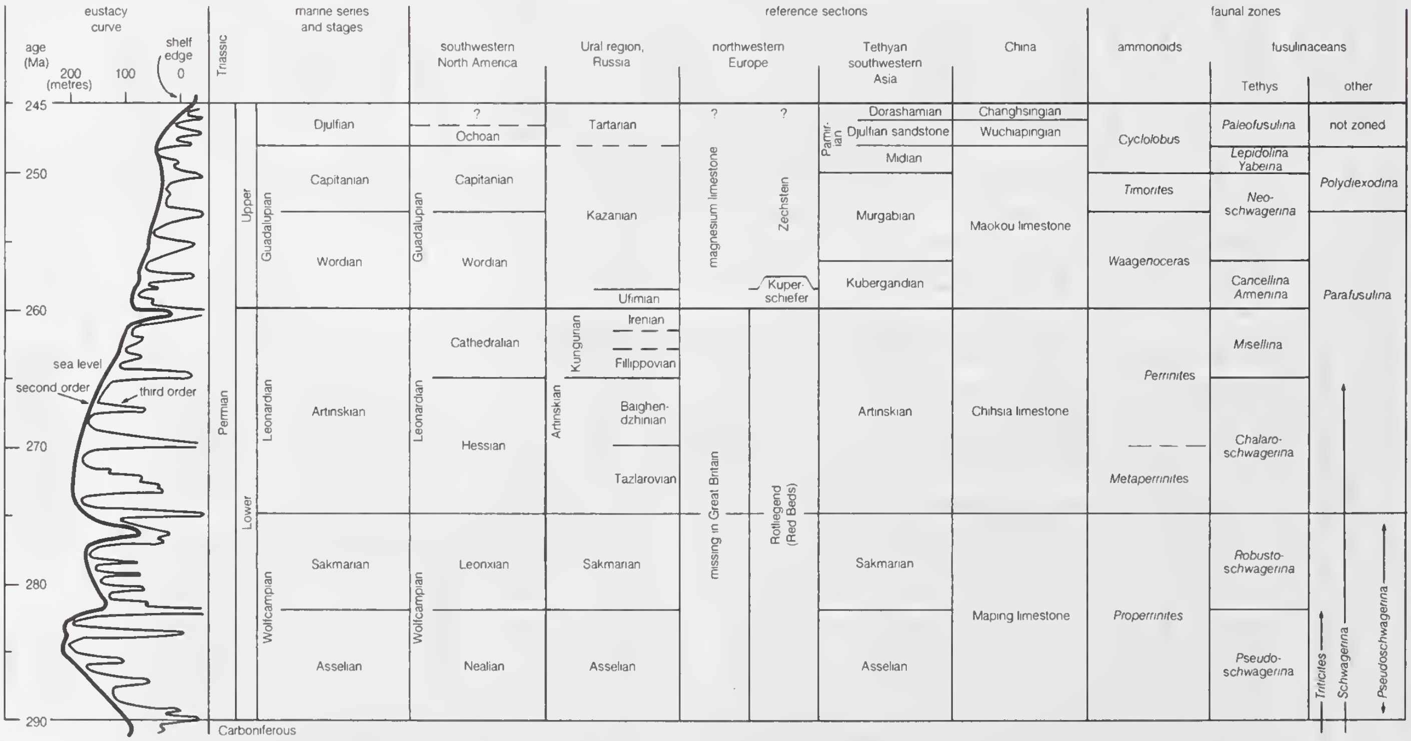
Problems of strongly differentiated faunal provinces during the Permian Period have plagued efforts to establish firm stratigraphic correlations between many regions. This differentiation of provincial faunas and their isolation from one another increases noticeably in the middle and later parts of the period. In the type area of the Permian, the Asselian, Sakmarian, and Artinskian stages are predominantly marine deposits with a reasonable number of cosmopolitan fossils. The Kungurian Stage is evaporitic and has locally limited and rare marine fossils. The Ufimian is largely nonmarine fluvial clastics, and the Kazanian is dolomitic and silty and has rare, restricted faunas that have been considered tolerant to wide swings in salinity—*i.e.,* from hyper- to hyposaline. The Tatarian is continental and has a well-described vertebrate, plant, insect, and freshwater ostracod fossil record.

In many other parts of the world, the equivalents of the Kungurian through Tatarian stages are marine deposits that have diverse and abundant normal marine faunas. Two predominantly carbonate provinces are recognized. One includes the southwestern United States and northwestern South America. The other, which is much larger and has a more diverse fauna, includes a belt of rocks from Tunisia and the Carnic Alps on the west through Turkey, Iran, southern China, Southeast Asia, and Japan, and central British Columbia and Washington, Oregon, and California in North America. This second carbonate province, called the Permian Tethys, was thoroughly disrupted by post-Permian orogenic deformation (as the result of seafloor spreading and plate tectonics) and is now found only as geographically dislocated fragments. These Permian Tethyan strata stand in sharp contrast to the Permian beds in the type area, and some geologists prefer to divide them into three series—Lower, Middle, and Upper. The two predominantly carbonate provinces are considered to have been tropical and subtropical and to have been centred near the equator but on opposite sides of the huge supercontinent of Pangaea.

Permian Tethyan strata

The boundary between the Permian System and the overlying Triassic System nearly everywhere constitutes a hiatus of one to several million years, except in the central Tethys region where local deposition was apparently continuous. There, the boundary between these two important systems—indeed, the boundary between the Paleozoic and Mesozoic—is not easily defined. The latest Permian faunas were reduced to only a few remnant species that were obviously sensitive to stressful new environments. Typical Triassic lineages were just starting out and also were few in number and species. The two sets of lineages appear to overlap in this part of the succession, and the exact placement of the boundary is still under study.

Economic significance. Permian rocks have long been important economic sources of evaporite minerals, such as halite (rock salt), sylvite (potash salts), gypsum and anhydrite, petroleum, and coal. The distribution of these



C A Ross and J R P Ross in C.A Ross and D Haman, (eds) Cushman Foundation for Foraminiteral Research, Special Publication 24, 1987

Figure 31: The marine series and stage names of the Permian, the correlation of regional reference sections, second and third order eustacy curves, and a broad, generic-level biostratigraphic zonation of ammonoids and fusulinaceans.

resources, in part, is related to the Permian paleolatitudes of the deposits (Figure 32). The evaporites were particularly common in subtropical and tropical Permian paleolatitudes in what is now West Texas, New Mexico, and Kansas in North America as well as in northwestern Europe and the European sector of Russia. Thick coals formed in cool temperate paleolatitudes, such as central and northern Siberia, Manchuria, Korea, peninsular India, eastern Australia, South Africa, Zimbabwe, and the Congo, which were all in high-paleolatitudes during the Permian (see Figure 32).

Many of the Permian intracratonic marine basins are productive sources of petroleum and have good reservoirs. The most famous are in West Texas, New Mexico, and Oklahoma in the United States and along the Ural fold belts in Russia.

Phosphorites are common in the deep-water sedimentary wedges next to the shelf margin in Montana, Idaho, Wyoming, Utah, and Nevada that marked the western edge of the North American craton in Permian time. In Europe, phosphorites occur along a deep-water trough marking the eastern edge of the Russian Platform.

Adapted from C A Ross and J R P Ross, Cushman Foundation for Foraminiferal Research, Special Publication 24

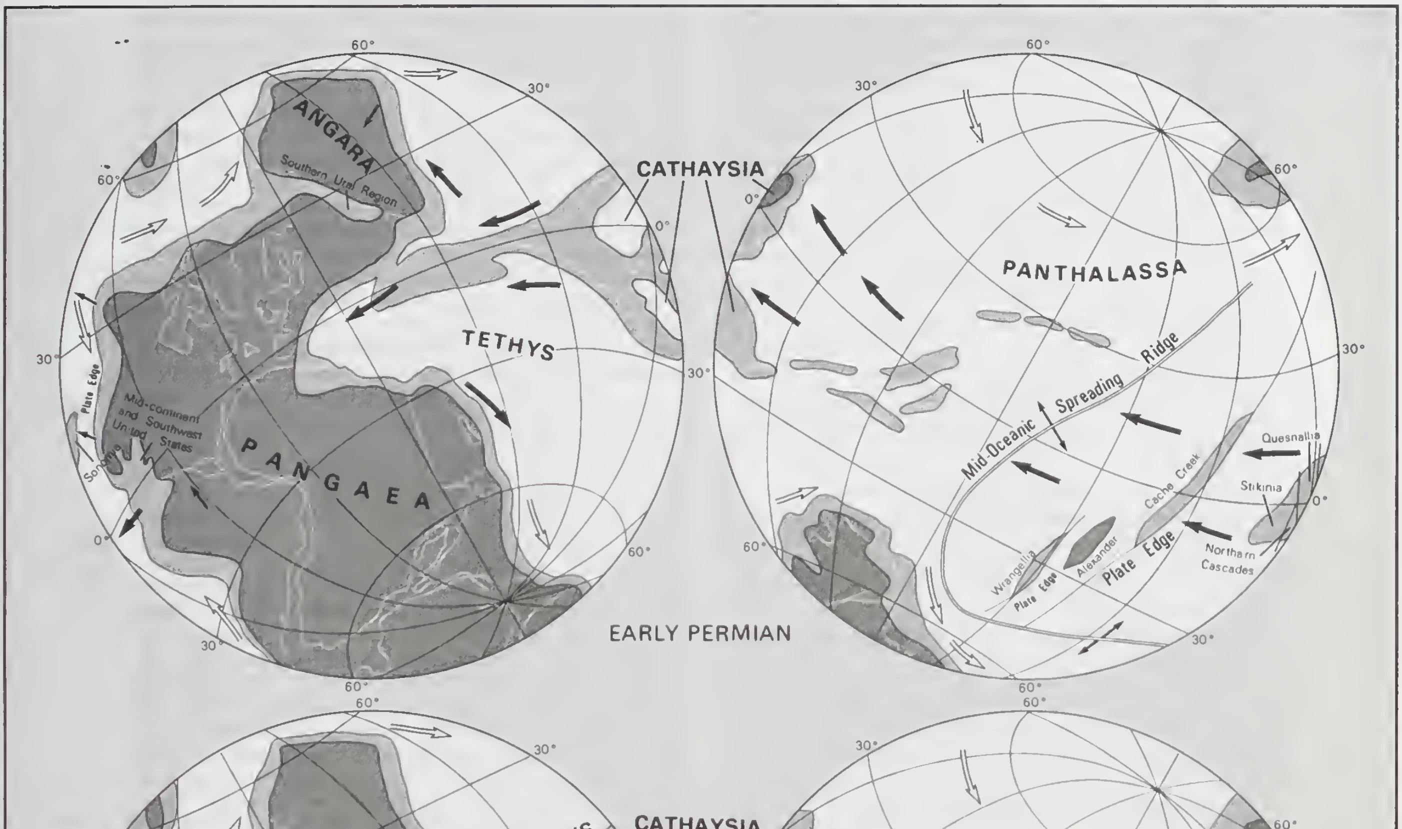


Figure 32: Paleogeography and paleoceanography of (top) the Early Permian and (bottom) the early Late Permian.

Of significance to European civilizations is the Permian Kupferschiefer, a copper-bearing shale that has been mined for hundreds, perhaps even thousands, of years. This rare type of copper ore was deposited in a marine basin that had no oxygen in its bottom waters.

Permian rocks. *Occurrence.* Permian rocks are found on all present-day continents; however, some have been displaced considerable distances, sometimes thousands of kilometres from their original site of deposition (Figure 32) by subsequent tectonic transport during the Mesozoic and Cenozoic eras. For example, Permian glacial and glacial marine deposits typical of the cold high latitudes of the Southern Hemisphere during the Permian are now found in Antarctica, southern Africa, India, Thailand, and Tibet, while Permian glacial deposits of the Northern Hemisphere are found in northeastern Siberia. By contrast, some Permian tropical and subtropical carbonate deposits, typical of deposition in low latitudes, have been relocated to high latitudes, and the present location of certain tropical provincial faunas suggests that other deposits have been moved considerable distances longitudinally to form tectonic belts of accretionary material added during Mesozoic and Cenozoic times to the former late Paleozoic continental margins. The commonly conflicting occurrences of different faunal provinces and tropical deposits in juxtaposition with temperate and cold (glacial) deposits became more readily explainable once the rates of motion that result from seafloor spreading and plate tectonics became established as 0.5 to 1.5 centimetres per year. When viewed on a Permian paleogeographic reconstruction (Figure 32), these apparent depositional conflicts disappear and a climatically compatible depositional pattern emerges.

Glacial and glacial marine deposits

The principal geographic features of the Permian world were the supercontinent Pangaea (which included all the then-existing major continents except North and South China) and a huge ocean basin called Panthalassa, with its branch, the Tethys (a large indentation in the tropical eastern side of Pangaea).

Panthalassa and Tethys encompassed scattered fragments of continental crust (microcontinents) and basaltic volcanic island arcs that featured extensive fringing limestone reefs and platforms. They are generally viewed as being analogous to the present-day Indian and Pacific ocean basins in terms of geologic construction. Cathyasia, comprising both the continents of North and South China, lay within western Panthalassa. The existence of several other, smaller (or now disrupted) microcontinents also has been proposed.

Pangaea contained extensive high mountain ranges along its orogenic suture between Euramerica (also known as Laurussia in Devonian and Carboniferous times) and the South American–Northwest African portion of Gondwana. These mountains influenced local climates and sedimentation during the early part of the Permian to a considerable extent. Later, during the middle Permian, the Angaran portion of western Siberia joined eastern Euramerica to form the Ural orogenic belt and mountains. Parts of Pangaea continued to be sheared and deformed by large linear zones of tear faults (steeply inclined faults along which movement has been largely horizontal and which include several failed rift structures) and by the vertical warping and faulting of the persistent North American transcontinental arch and similar tectonic activity in the northwestern segment of Europe.

Because these orogenic and related tectonic events progressively changed sedimentation conditions during the Permian Period, it is possible to consider shifts in depositional patterns in several phases. The Asselian and Sakmarian represent intervals of gradual transition from Late Carboniferous depositional facies typified by large sea-level fluctuations and strongly developed cyclical sediments (Figure 31). Less rapid and pronounced extremes in sea-level changes and more uniform sedimentation characterized the Artinskian. The Kungurian and later Permian deposition reflect increasing exposure and aridity on the cratonic shelves and well-developed intracratonic basins with more extensive evaporites. Marine deposits were primarily along the shelf margins. The latest phase, seen in Djulfian sedimentation, is not widely distributed and probably was restricted to the cratonic shelf margins or even to the upper shelf slopes.

Impact of tectonic activity on depositional patterns

Distribution and types. In terms of geologic setting, the Permian sediments that were deposited as thick sedimentary wedges along the tectonically active margins of the major cratons are the least understood in detail. Because there was one megacontinent, Pangaea, most of these wedges were facing into the huge active ocean basin of Panthalassa–Tethys. All these Permian sediments have subsequently been thrusted, metamorphosed, and involved in major geologic deformation. In many cases, it is difficult to determine the original thickness of these wedges, and much of the fossil evidence is from clastic material that was derived from shallow shelf environments or eroded from older rocks and deposited as deepwater debris fans. Great deposits—perhaps originally one to three kilometres thick—are known in central Nevada, Idaho, and northward into Canada. Similar deposits have been found in the Middle East, China, Japan, and eastern Siberia.

Interleaved with these thick clastic wedges are other thrust slices of ocean-floor deposits. These are thinner, about 0.5 kilometre thick or less, and are characterized by radiolarian-rich cherts, basaltic volcanic dikes, sills, and submarine lava flows, as well as silts and clays (commonly metamorphosed to slates) of the distal ends of turbidity flows (*i.e.,* those of density currents). Such completely deformed deposits are the only remaining record of Permian (and older) ocean-floor deposits, because only Jurassic and younger oceanic sediments survive in the present-day ocean basins. Most of the Permian (and older) ocean-floor deposits and thick sedimentary wedges have been caught up in subduction zones along the plate boundaries and lost to the Earth's mantle.

Associated with some oceanic basalts are thick accumulations of reef limestone that formed on seamounts and volcanic island arcs. These reefs were tropical and subtropical, as are modern oceanic limestone reefs. Because limestone is comparatively less dense than adjacent oceanic rocks, such as basalt or chert, many of the Permian reef limestones were not as easily subducted and are present in many modern mountain belts. Limestones of this kind are known from Tunisia, the Balkan Peninsula, Turkey, the Crimea (in Ukraine), the Middle East, northern India, Pakistan, Southeast Asia, New Zealand, China, Japan, eastern Siberia, Alaska, and the western Cordillera of Canada, the United States, and northwesternmost Mexico. These limestones commonly have complex structural relationships with adjacent rocks, usually of Triassic, Jurassic, or early Cretaceous age, suggesting that most such Permian limestones were incorporated into the mountain belts by those times.

Reef limestone

Cratonic shelf sedimentation in low paleolatitudes during the Permian was characterized by the gradual withdrawal of shorelines and the progressive increase in eolian sands, red beds, and evaporites. Many intracratonic basins, such as the Anadarko, Delaware, and Midland basins in the western United States, the Zechstein basin of northwestern Europe, and the Kazan Basin of eastern Europe, show similar general changes. In most cases, the inner parts of these basin systems became sites of red bed deposition during the Leonardian, followed within a short time by locally extensive evaporites. Sand sources along the Ancestral Rocky Mountains supplied eolian sand and silt in great quantities.

The outer portions of the intracratonic basin systems, as in the Delaware and Zechstein basins, were involved in some transform faulting and extensional tectonics and locally developed considerable depositional relief. Although some of this relief was from rotated fault blocks, most of it resulted from the very rapid growth of limestone reefs on upthrown blocks (*i.e.,* the sides of faults that appear to have moved upward) and the slower accumulation of clastic sediments on downthrown blocks. Striking examples of these reefs form the Guadalupe Mountains of West Texas and New Mexico. Such reefs also occur in the subsurface along the Central Basin Platform in West Texas, where they produce petroleum. Similar reefs are known from northern England, Germany, and the subsurface of the

North Sea. Lower Permian limestone reefs also are known from the western and southern Urals of eastern Europe.

At higher paleolatitudes, limestone is rare, and clastic rocks dominate the succession. Australia, Namibia, South Africa, peninsular India, southern Tibet, and southern Thailand all report Permo-Carboniferous tillite. These areas, as their paleogeographic reconstruction indicates, would have been in relatively high latitudes as parts of Gondwana during the Permian. Tillites are also known from the northern high paleolatitudes in northeastern Siberia. Some of the Gondwanan areas were tectonically very active during Permian time, as evidenced by extensive basaltic, andesitic, and other volcanic rocks in eastern Australia and by the development of intracratonic sedimentary marine basins, such as the Carnarvon Basin in Western Australia where nearly five kilometres of Permian sediments accumulated.

Continental rocks were widespread on all the cratons during the Permian. The Dunkard Group is a limnic coal-bearing succession that was deposited from latest Carboniferous into Early Permian time along the western side of the then-newly formed Appalachian Mountains. In what is now Kansas and areas of the western United States, coal-bearing cyclothems that formed along the margins of shallow seas in continental interiors persisted well into the Early Permian. Coal-bearing Lower and Upper Permian beds, up to three kilometres thick, are widely distributed in Australia, peninsular India (the lower part of the Gondwana System), southern Africa (the lower part of the Karoo [also spelled Karroo] System), the Kuznetsk Basin of southwestern Siberia, and the Paraná and Precordillera basins of southern Brazil and western Argentina in South America. Red beds were common in continental beds in tropical and subtropical paleolatitudes. During the Permian Period red beds became more abundant and widespread, suggesting that the climate became progressively warmer and tropical conditions became more widespread. This warming trend is also hinted at by a significant increase in the amount of dolomite in shelf sediments.

Permian environment. *Paleogeography.* The Permian constitutes an important crossroads both in the history of the Earth's continents and in the evolution of terrestrial life. During roughly the first half of the period, Gondwana collided with and joined western Euramerica, to which the Angaran sector of Siberia was subsequently fused (see Figure 32). Thus, the assembly of what is often referred to as Greater Pangaea was completed by mid-Permian time, giving rise to a single mountainous continental landmass that extended across all the climatic temperature zones without interruption virtually from one pole to the other. This megacontinent was surrounded by the immense world ocean Panthalassa, which, with the Tethys Sea, was the site of a small number of microcontinents, island arcs, oceanic plateaus, and trenches.

Formation of Pangaea

Paleoclimate. Extensive glaciation persisted in the Early Permian, largely in what is now India, Australia, and Antarctica but also in Siberia near the north paleopole. Hot, dry conditions prevailed elsewhere on Pangaea, and deserts became widespread in various tropical and subtropical areas of the continent by the Late Permian.

The gradual climatic warming that took place during the Permian at first encouraged evolutionary expansion among shallow-water marine faunas but later resulted in marked extinctions (see below *Permian life*). On the other hand, this warming trend, combined with climate diversity, provided an opportunity for broad adaptive radiation in terrestrial plants, insects, and reptiles, particularly among mammallike reptiles.

Permian life. Life during the Permian was very diverse, and the marine life of the period was perhaps more diverse than that of modern times. The previous period, the Carboniferous, had had two instances of significant marine faunal extinctions that had followed one another in relatively rapid succession. One was at the end of the Early Carboniferous (Mississippian subperiod in North America) and the other at the end of the Middle Carboniferous (the close of the Middle Pennsylvanian). Both may be attributed to global cooling during continental glaciations. The latest Carboniferous witnessed the establishment of new or highly modified marine lineages with relatively little ecological competition from the remnants of what had been highly successful earlier phylogenetic lineages. Most of these lineages are identified as new families or suborders among the foraminifers, ammonoids, brachiopods, bryozoans, bivalves, and some less-studied groups. In Wolfcampian, Leonardian, and Guadalupian times, these newly established lineages underwent rapid evolution and filled a remarkable number of specialized marine ecological niches. Some of this rapid diversification was probably the result of filling the ecological niches left vacant by the extinctions of the Carboniferous. Other factors undoubtedly included the gradual warming of oceans during the latest Carboniferous and the more rapid warming trends of the Early Permian. In paleotropical areas, a succession of carbonate bank, bioherm, and reef faunal associations evolved, which culminated in the late Guadalupian Capitan Reefs of the western United States and the more faunally diverse associations that formed even thicker fringing reefs in the Tethyan area. Because these two sets of tropical marine shallow-water faunal associations were separated by the large supercontinent Pangaea on the one hand and the deep oceanic basin of Panthalassa on the other, they tended to evolve independently of each other.

Near the end of the Guadalupian both of the these tropical faunal realms suffered major but incomplete extinctions, and the Djulfian saw a brief and relatively minor evolutionary re-expansion in some of the foraminifers and ammonoids. The trilobites were extinct by the end of the Leonardian. Only a few Permian bryozoan and brachiopod genera, and only one or two species of those genera, survived into the earliest Triassic. Although the magnitude of the extinctions among shallow-water marine organisms during the later parts of the Permian was great, the process took several million years and was accomplished in a series of steps followed by unsuccessful attempts by the surviving faunas to rediversify.

Permian extinctions

Terrestrial life in the Permian was closely keyed to the evolution of terrestrial plants, which of course were the primary food source for terrestrial animals. The fossil plant record for the Permian consists predominantly of ferns, seed ferns, and lycophytes, which is attributable to their adaptation to marshes and swampy environments. A less abundant fossil record of early coniferophytes and even some protoangiosperms suggests a broad adaptation of these plant groups to progressively drier areas. As discussed earlier, evidence seems to point to gradually warming and drier climates, which would have encouraged plant adaptations to drier conditions.

Another line of evidence suggesting broad plant diversification is found in the evolution of insects; these animals tend to be highly selective in choosing their plant hosts. Among the superclass Hexapoda of the phylum Arthropoda, at least 23 orders are known from the Permian, and of these orders 11 are extinct. By comparison, 250 million years later, there exist only 28 insect orders, and the new orders are mainly those that have adapted to living on the angiosperms or mammals that evolved after Permian time. Permian insects include a huge dragonfly-like creature that had a wingspan of 75 centimetres.

Terrestrial vertebrates of the Permian, in addition to freshwater sharks and fish and several orders of relatively large amphibians, are noted for the first appearance of several important reptile lineages. Although a few primitive and generalized reptile fossils are found in Middle Carboniferous deposits, Permian reptile fossils are locally common and include the protorosaurs, aquatic reptiles; the captorhinomorphs, the "stem reptiles" from which most other reptiles are thought to have evolved; the eosuchians, early ancestors of the snakes and lizards; early anapsids, ancestors of turtles; early archosaurs, ancestors of the large ruling reptiles of the Mesozoic; and the synapsids, a common and varied group of mammallike reptiles that eventually gave rise to mammals in the Mesozoic. Of these, the captorhinomorphs and synapsids are probably the best known.

Emergence of important reptiles

Captorhinomorphs are common in the Lower Permian beds of North America and Europe. Massively built and large for their day, they reached lengths of two to three metres. Captorhinomorphs are less common in Upper Permian beds, and only one small group survived into the Triassic.

Synapsids are divided into two orders: the pelycosaurs and the therapsids. The Early Permian pelycosaurs included a lineage containing both carnivorous and herbivorous members that developed long spines on their vertebrae, which seem to have supported a membrane, or "sail." The function of the sail is not fully understood; however, suggestions include its use in regulating body temperature. Pelycosaurs reached 3.5 metres in length and had large teeth. Their remains are commonly found in the Lower Permian red beds of central Texas but are rare in Europe.

The therapsids were advanced synapsids that are known from the Upper Permian Karoo beds of South Africa, South America, and India and equivalent beds in Scotland and what was formerly the Soviet Union. Therapsids range into the Triassic and show a great deal of diversification. Their dentition and bone structure are remarkably mammallike, and the point at which a mammallike reptile passes into an actual mammal has long been a point of controversy. The success of therapsids in the relatively high paleolatitudes of Gondwana has strengthened the view that they were able to maintain an elevated body temperature. (J.R.P.R./Ch.A.R.)

Mesozoic Era

The Mesozoic (from the Greek for middle life) began about 245 million years ago and ended 66.4 million years ago (see Table 4). The major divisions of the era, from oldest to youngest, are the Triassic, Jurassic, and Cretaceous periods. The Mesozoic was a time of heightened tectonic activity during which the supercontinent of Pangaea fragmented into separate continents that were gradually scattered across the Earth in a nearly modern geographic distribution. It also was a time marked by a distinct modernization of life-forms; the ancestors of the major plant and animal groups that exist today first made their appearance.

Redistribution of the landmasses

At the outset of the Mesozoic, all the continents were still joined together. Continental rifting, however, began in Late Triassic to Early Jurassic times. The separation of Laurasia and Gondwana and their constituent continents started by the Middle Jurassic, while much of Pangaea lay between 60° N and 60° S. (The paleoequator cut through the widening Tethys seaway between Laurasia and Gondwana.) Spreading centres and mid-oceanic rifts formed between several of the separating continents as well as between the segments of Gondwana. During the Jurassic, North America began pulling apart from Eurasia and Gondwana. By the Late Jurassic, Africa had started to split off from South America, and Australia and Antarctica had separated from India. Near the close of the Cretaceous, Madagascar separated from Africa, and South America drifted northwestward.

Thick sequences of marine sediments accumulated in large linear troughs called geosynclines along passive continental margins—namely, those formed by continental rifting and rupture. Geosynclinal deposits of Jurassic age formed in the present-day circum-Pacific region, along the coasts of eastern North America and the Gulf of Mexico, and on the margins of Eurasia and Gondwana (*i.e.*, along the northern and southern boundaries of the Tethys seaway). Thick deposits of this Tethyan Geosyncline were incorporated into the Alpine-Himalayan mountain system beginning in late Mesozoic time.

Major mountain building began on the western margin of both North America and South America and between the fragments of Gondwana. For example, the northwesterly movement of North America resulted in the collision of the western edge of the North American continental plate with a complex of island arcs in the Late Jurassic. During the ensuing mountain-building episode known as the Nevadan orogeny, so-called suspect, or exotic, terranes (geologic provinces that originated in the ocean crust and differ markedly in stratigraphy, paleomagnetism, and paleontology from the adjoining continental crust) were accreted to the margin of the North American Plate, huge granitic batholiths formed in what is now the Sierra Nevada range along the California-Nevada border, and thrusting occurred in an eastward direction. Other notable Mesozoic episodes of mountain building include the Sevier and Laramide orogenies that took place in western North America during Cretaceous time.

Episodic marine transgressions and regressions

At various times during the Mesozoic, shallow seas invaded continental interiors and then drained away. During mid-Triassic time, a marine incursion—the Muschelkalk Sea—covered the continental interior of Europe. Seas again transgressed between the Early and Late Jurassic and in the Early Cretaceous. Marine waters flooded large segments of all the continents during mid-Cretaceous time, marking the last transgression of a global scale. The sharp rise in sea level and resultant worldwide flooding are thought to have been caused chiefly by accelerated seafloor spreading and the attendant enlargement of the ocean ridges. This displaced enormous amounts of ocean water onto the landmasses. Transgression was so extensive that in North America, for example, a shallow sea spread all the way from the Arctic to the Gulf of Mexico.

Mesozoic rocks are distributed widely, appearing in various parts of the world. A large percentage of these rocks are of the sedimentary variety. Triassic limestone formed largely in the Tethyan (present-day Mediterranean) region and the western United States, while graywackes, sandstones, and shale of the same age predominate in the circum-Pacific geosynclines. Extensive beds of limestone, sandstone, ironstone, and clays were left in continental interiors by the shallow epicontinental seas of the Jurassic. Likewise, the large-scale marine inundation of Cretaceous time resulted in the widespread deposition of chalk, clay, and marl in the lower areas of several continents.

A substantial amount of igneous rock also formed during the Mesozoic. The initial rifting of Pangaea produced fault-block basins that were later filled with basaltic intrusions and pillow lavas, as well as with fluvial and lacustrine beds, between Late Triassic and Early Jurassic time. Subsequent rifting of the Gondwanan segments was accompanied by outpourings of flood basalt (*i.e.*, plateaus of basalt that extend many kilometres in flat, layered flows). The orogenies of the middle and late Mesozoic involved volcanism and plutonic intrusion, as with the emplacement of granitic and andesitic plutons in the Andes of South America during the Late Jurassic.

Mesozoic fauna and flora

Mesozoic biota had to recover from the major extinction episode at the end of the Paleozoic Era. Vertebrates, which appear to have been less severely affected by the event than invertebrates, diversified progressively through the Triassic. During this time, the seas became inhabited by such marine reptiles as the nothosaurs and ichthyosaurs. The Triassic terrestrial environment was dominated by the therapsids, mammallike reptiles, and the thecodonts, ancestors of the dinosaurs of the later Mesozoic. The first true mammals, small shrewlike omnivores, appeared in the Late Triassic, as did the primitive forebears of lizards, turtles, and crocodiles. A second major extinction event struck at the close of the Triassic, one that wiped out as many as 400 genera of ammonoids, along with about 80 percent of the reptiles. In all, 35 percent of the then-existing animal groups suffered extinction.

During the Jurassic, the ammonoids and brachiopods (lamp shells) recovered from the Late Triassic crisis, thriving in the warm epicontinental seas. The ammonoids, in fact, rapidly became the most prominent invertebrate marine form, and their remains proved to be the most important index fossil for worldwide correlation of Jurassic rock strata. Many other animal forms, including mollusks (notably the bivalves), sharks, and bony fish, flourished during the Jurassic, and new forms, such as frogs, toads, and salamanders, emerged. All these groups, however, were overshadowed by the giant reptiles that reigned throughout much of the Jurassic and Cretaceous. Such dinosaurs as the fierce predator *Tyrannosaurus* and the enormous vegetarian *Apatosaurus* dominated on land. Other reptile forms, like the plesiosaurs, ruled the seas, while the pterosaurs, creatures with batlike membranous wings, dominated the sky.

Birds evolved from reptilian ancestors during the Late Jurassic, and two important modern mammal groups—the placentals and the marsupials—made their appearance in the Late Cretaceous. Plant life also exhibited a gradual change toward more modern forms during the course of the Mesozoic. Whereas seed-fern flora had predominated in the Triassic, forests of palmlike gymnosperms known as cycads and conifers proliferated under the tropical and temperate conditions that prevailed during the Jurassic. The first flowering plants, or angiosperms, appeared by the Cretaceous. They radiated rapidly, supplanting many of the primitive plant groups to become the dominant flora form by the end of the Mesozoic.

Decimation of many animal forms

The Mesozoic closed with an extinction event that devastated many forms of life. The ammonoids, reef-building rudist bivalves, and numerous varieties of foraminiferans and coccolithophores died off, as did the dinosaurs and flying and marine reptiles. The Late Cretaceous extinctions have been variously attributed to such phenomena as global tectonics, draining of the epicontinental seas, northward migration of the continents into different climatic zones with much cooler conditions, intensified volcanic activity, and a catastrophic meteorite impact. The extinctions, however, were not sudden; they spanned millions of years, a fact that has led some researchers to believe that, after an era of dominance among various forms of life, decay of the Mesozoic ecosystem had set in. The Cretaceous extinction event may very well have had multiple causes. As the landmasses were uplifted by plate tectonism and migrated poleward, the climate of the Late Cretaceous began to change from warm/moist and warm/dry to cool. This favoured the angiosperms over tropical vegetation and provided habitats and food sources for the fur-covered mammals. In effect, conditions were right for the life-forms that were to replace the dinosaurs. (L.Si.)

TRIASSIC PERIOD

General considerations. The Triassic, the first period of the Mesozoic Era, began about 245 million years ago at the close of the Permian Period and ended approximately 208 million years ago when it was succeeded by the Jurassic Period (see Table 4). The rocks that originated during this time interval make up the Triassic System, which consists mostly of rocks of sedimentary type. Because there are relatively few igneous rocks to provide reliable radiometric dates, the time span and absolute ages cited by different investigators for the Triassic Period tend to vary (*e.g.,* some indicate that the period extended from about 240 million to 195 or 200 million years ago). Such dates are subject to revision as new and more accurate age determinations are made.

The name Trias (later modified to Triassic) was first proposed in 1834 by the paleontologist Friedrich August von Alberti for a sequence of strata in central Germany comprising three formations: the Bunter (or Buntsandstein), Muschelkalk, and Keuper. This so-called Germanic facies, typical of the Triassic in northern Europe, has many drawbacks as a standard by which to correlate rocks from other regions, because the Bunter and Keuper are mostly nonmarine, while the Muschelkalk, though marine, is not widespread geographically and contains an abnormal marine fauna. For many years the traditional Triassic stages were based mainly on type sections from the marine Triassic of the Alps, but now more complete marine sequences have been discovered in North America to serve as a standard for Triassic time in general. Also, seafloor spreading and plate tectonics have yielded important new information on the paleogeography and paleoclimatology of the Triassic, allowing for a better understanding of the evolution, extinction, paleoecology, and paleobiogeography of the biota. At the same time, the problems of defining the lower and upper boundaries of the Triassic System on a worldwide basis and understanding the reasons for mass extinctions at these boundaries continue to occupy the attention of paleontologists.

Major subdivisions. The Triassic Period is divided into three chronostratigraphic epochs: Early, Middle, and Late. In like manner, the Triassic System is broken down into three lithostratigraphic series: Lower, Middle, and Upper. Each of these series is further subdivided into stages and biochronological zones (or biozones), based mainly on the precise vertical ranges of marine pelagic cephalopod mollusks called ammonoids.

Series and stages

Early subdivision of the Triassic into epochs, stages, and biozones was based primarily on the extensive and highly fossiliferous Alpine (western Tethyan) sequence of marine strata exposed in Austria, Italy, Germany, and Switzerland. It was here that the type sections, or stratotypes, for the Middle Triassic stages Anisian and Ladinian and the Upper Triassic Stages Carnian, Norian, and Rhaetian were first established. The Scythian Stage of the Lower Triassic was based on occurrences in the Himalayas, the Salt Range of Pakistan, and what was once the Soviet Union but has now been replaced by three new stage names, which, in ascending order, are Griesbachian (with stratotypes in the Himalayas [substage Gangetian] and Arctic Canada [substage Ellesmerian]; Nammalian (with stratotypes in Arctic Canada [substages Dienerian and Smithian]); and Spathian (with stratotype in Arctic Canada). It should be noted, however, that the above stage and substage names have not been universally accepted; alternative names for part or all of the Triassic are used in Russia (as well as in the other former Soviet republics), Japan, and New Zealand. The complete sequence of epochs, series, stages, and biozones recognized by most authorities is shown in Table 18.

Distinctive features. The Triassic Period marked the beginning of the major changes that were to take place throughout the Mesozoic in the distribution of the continents and the impact of those changes on paleoclimates, paleobiogeography, and the evolution of Triassic biotas. Shallow shelf seas were reduced in extent toward the end of the Permian, probably contributing to the extinction of several marine invertebrate groups, such as trilobites, rugose and tabulate corals, productacean brachiopods, cryptostomate and fenestrate bryozoans, blastoid echinoderms, and fusulinid foraminiferans. The supercontinent of Pangaea comprised virtually all the major landmasses of the world at the beginning of the Triassic. Terrestrial environments with warm, dry climates predominated, which resulted in an increase in the relative importance of land animals, including the first mammals at the end of the Triassic. Reptiles increased in diversity and number, heralding the great radiation that would characterize this group during the Jurassic and Cretaceous periods. Some groups of marine invertebrates, such as the ammonoids, made a dramatic recovery from near-extinction in the Late Permian, although they were to suffer yet another crisis in the Late Triassic. Shelf seas, which gradually became more extensive during the Middle and Late Triassic, were also colonized for the first time by large marine reptiles and reef-building scleractinian corals of modern aspect, while holdovers from the late Paleozoic included bivalves, brachiopods, echinoids, and gastropods. In summary, the Triassic ushered in great changes not only in the distribution of the continents but also in the evolution of living forms on land and in the oceans.

Economic significance. Few deposits of major economic importance were formed during the Triassic. Workable coal deposits are known from Arctic Canada, Russia, Ukraine, China, Japan, Australia, and Antarctica. Oil and gas occurrences are rare, however. Halite (rock salt) is mined from Triassic evaporites in England, France, Germany, and Austria.

Development of geosynclines around Panthalassa

Triassic rocks. *Occurrence and distribution.* Major depositional troughs—*i.e.,* geosynclines—developed around Panthalassa, the ancestral Pacific Ocean, during the Early and Middle Triassic. Great quantities of marine sediments, mainly sandstones, shales, and graywackes, collected in these troughs, as indicated by deposits now found in the western Pacific geosynclinal belt (New Zealand and Japan) and the eastern geosynclinal belt (Alaska, Arctic Canada, British Columbia, western United States, and the west coast of South America). As an example, more than 3,000 metres of Triassic sediments accumulated in the Sverdrup Basin of Arctic Canada. A deep, narrow arm of Panthalassa, the Tethys Sea, stretched along an east–west belt separating what is now Africa from south-

Table 18: Series, Stages, Substages, and Biozones of the Triassic System*

series	stage	substage	zones (stratotype† in Tethys, others in North America)		
Upper Triassic	Rhaetian		*Choristoceras crickmayi*		*Choristoceras marshi*†
	Norian	Upper Norian	*Cochloceras amoenum*		
			Gnomohalorites cordilleranus		
		Middle Norian (Alaunian)	*Himavatites columbianus*		*Halorites macer*†
					Himavatites hogarti†
			Drepanites rutherfordi		*Cyrtopleurites bicrenatus*†
		Lower Norian	*Juvavites magnus*		
			Malayites dawsoni		*Malayites paulckei*†
			Stikinoceras kerri		*Guembelites jandianus*†
	Carnian	Upper Carnian (Tuvalian)	*Klamathites macrolobatus*		*Anatropites beds*†
			Tropites welleri		
			Tropites dilleri		*Tropites subbullatus*†
		Lower Carnian (Julian)	*Austrotrachyceras obesum*		*Austrotrachyceras austriacum*†
			Trachyceras desatoyense		*Trachyceras aonoides*†
Middle Triassic	Ladinian		*Frankites sutherlandi*		
			Maclearnoceras maclearni		*Protrachyceras archelaus*†
			Meginoceras meginae		
			Progonoceratites poseidon		
			Eoprotrachyceras subasperum		*Eoprotrachyceras curionii*†
	Anisian	Upper Anisian (Illyrian)	*Frechites chischa*	*Frechites occidentalis*	
				Parafrechites meeki	*Ticinites polymorphus*†
			Frechites deleeni	*Gymnotoceras rotelliformis*	*Paraceratites trinodosus*†
		Middle Anisian (Pelsonian)		*Balatonites shoshonensis*	*"Paraceratites" binodosus*†
			Anagymnotoceras varium		*Anagymnotoceras ismidicum*†
				Acrochordiceras hyatti	*Nicomedites osmani*†
		Lower Anisian (Aegean)	*Lenotropites caurus*		
Lower Triassic (= Scythian)	Spathian		*Keyserlingites subrobustus*	*Neopopanoceras haugi*	
				Subcolumbites beds	*Tirolites cassianus*†
			"Olenikites" pilaticus	*Columbites parisianus*	
	Nammalian	Smithian	*Wasatchites tardus*		*Anasibirites pluriformis*†
			Euflemingites romunderi		*Hedenstroemia himalayica*†
		Dienerian	*Vavilovites sverdrupi*		*Gyronites frequens*†
			Proptychites candidus		
	Griesbachian	Upper Griesbachian (Ellesmerian)	*Proptychites strigatus*		
			Ophiceras commune		*Ophiceras connectens*†
		Lower Griesbachian (Gangetian)	*Otoceras boreale*		
			Otoceras concavum		*Otoceras woodwardi*†

*Based on ammonoids.

Source: Modified from E.T. Tozer, *The Trias and Its Ammonoids: The Evolution of a Time Scale* (1984), Geological Survey of Canada, Miscellaneous Report 35

ern Europe, and it also received geosynclinal deposits. In the northern Tethyan geosyncline, these deposits now outcrop in the Alps, Turkey, Iran, Pakistan, and the Himalayas, mainly as limestones, with deep-sea sediments such as radiolarian cherts that formed in troughs in the deeper parts of the Tethys Sea. To the south was the southern Tethyan geosyncline, bordering Gondwana and stretching from northern India through the Middle East to northern Africa. Shallow shelf-sea embayments of limited distribution occurred landward of these geosynclines and are represented mainly by limestones in low latitudes, as around the margins of the Tethys Sea. Such tropical and subtropical shelf seas were warm and often supported small reefs, the forerunners of the more extensive coral reefs of today. Although the Permo-Triassic extinction of rugose and tabulate corals resulted in an absence of Lower Triassic corals, small reeflike mounds of early Middle Triassic age were succeeded later in Middle Triassic times by more extensive reef complexes with some Permian biotic elements retained. Such reefs have been described from the Tirolian Alps of Austria and the Dolomites of Italy. Late Triassic reef complexes, more modern in aspect and dominated for the first time by scleractinian corals, occur as thick sequences in the Dachstein and Steinplatte regions of Austria and Germany, as well as in Iran and the Himalayas.

Reef complexes

In the circum-Pacific region some shelf-sea deposits, generally of clastic facies (sandstones and shales), occur in Western Australia, Siberia, and the circum-Arctic region, including Arctic Canada, Alaska, eastern Greenland, and Spitsbergen.

Continental sediments, dominated by red beds (*e.g.*, sandstones and shales of red colour) and evaporites, accumulated on land throughout the Triassic Period. The Bunter and Keuper Marl of Germany and the New Red Sandstone of Britain are examples of such red beds north of Tethys, while to the south are similar deposits in India, Australia, South Africa, and Antarctica. Although deposits of this kind usually indicate accumulation in arid regions, such as inland desert basins, sediments of fluvial or lacustrine origin may also be represented by these red beds. Large basins containing Triassic continental sediments occur in South America (Colombia, Venezuela, Brazil, Uruguay, Paraguay, and Argentina) and in western North America, particularly in Utah, Wyoming, Arizona, and Colorado. In eastern North America great thicknesses of sedimentary rocks of continental origin were deposited during the Late Triassic in a series of fault basins, of which the Newark Basin is probably the best known. Here sequences of continental red clastics with dinosaur tracks and mud cracks, along with black shales containing fossils of freshwater organisms, indicate a depositional environment of rivers draining into freshwater lakes in a generally arid or semiarid region, which from paleomagnetic evidence appears to have been located about 20° N of the paleoequator.

Triassic igneous rocks are not common, and reliable radiometric dates are available only from Upper Triassic rocks. Examples of extrusive basalt flows are known from Australia, South America, and eastern North America, while intrusive rocks include the well-known Palisades Sill of the Newark series. This 300-metre-thick diabase intru-

Igneous rocks of the Upper Triassic

sion has yielded a potassium-argon age of 193 million years. Lava flows in the Hartford Basin of Connecticut have been used to estimate the age of the Triassic-Jurassic boundary as between 184 and 195 million years, based on potassium-argon and argon-40-argon-39 geochronology.

Types. As a broad generalization, Triassic geosynclinal and shelf-sea deposits formed in low paleolatitudes are dominated by limestones, with minor amounts of clastics (*e.g.*, sandstone, shale, and graywacke), while high-paleolatitude depositional basins and the circum-Pacific geosyncline are dominated by clastics, with minor amounts of limestone. Volcanism is usually associated with faulting and is represented by basalt flows and diabase intrusions.

Correlation. The Triassic System is dominated by sedimentary rocks that generally do not yield reliable radiometric dates. Since Triassic igneous rocks are rare, relative ages of the sedimentary rocks, derived from superposition, lithology, and biochronology, must be used for correlation. Of these three tools, biochronology has proved to be the most accurate and reliable. Although palynology may prove to be useful for correlation of marine and nonmarine strata in the future, the most widely used fossils in biochronology are those of ammonoids. This is because pelagic swimming forms of this type fulfill the basic requirements for ideal zone fossils in being widespread geographically, rapidly evolving, and substrate-independent. Ammonoids thrived in Triassic seas, along with pelagic bivalves, such as *Claraia* and *Halobia.* While ammonoids have been used successfully to erect a series of biozones, each one probably representing no more than one million years, the problem has been to find complete sequences of undisturbed marine strata that represent all stages of Triassic time in any one general region. Since the Germanic facies is mostly continental, the marine Triassic of the Alps has traditionally been used as a standard for the period, with the two most important localities being Salzkammergut in the northern Austrian Alps and St. Cassian (now San Cassiano) in the Dolomites to the south. Unfortunately, there are very few ammonoids common to both the Salzkammergut and St. Cassian sections. Indeed, the Alpine succession in general is not without its drawbacks when an attempt is made to determine sequential faunal relationships. In the red Hallstatt limestone facies in the Alps and throughout the Tethyan region, ammonoids often occur in lenses (*i.e.*, deposits bounded by converging surfaces that are thick in the middle and thin out toward the edges) in areas of tectonic complexity. Furthermore, faunas are often condensed through possible postdepositional submarine solution, resulting in "cemeteries" of ammonoids of different ages in close association. Also, fracturing and solution occurring at nearly the same time during the Triassic apparently caused local mixing and inversion of zones as younger beds collapsed into solutional voids in older strata. Such condensed and mixed assemblages have led to difficulties for paleontologists attempting to use the Alpine zonal scheme as a standard with which to correlate marine Triassic sequences in other regions. Nevertheless, the importance of the Alpine Triassic should not be underestimated in the history of Triassic studies, because its fossils permitted initial correlations to be made with the Germanic Muschelkalk and marine sequences in the Arctic, Pacific, Himalayas, and Salt Range of Pakistan by the end of the 19th century.

Reliance on fossil ammonoids for correlation

Isolated occurrences of marine Triassic rocks were known from western North America by 1890, but discoveries of several hundred new localities from this region and Arctic Canada between about 1955 and 1970 added much information on the biochronology of the region. It also was recognized that more than half the world's known genera of ammonoids occur in North America, testifying to the cosmopolitan nature of the group. Dissatisfaction with the problems of using the Alpine succession as a standard for Triassic time led to the proposal of a new zonal scheme based on relatively complete and in-place sequences in Arctic Canada, northeastern British Columbia, and the western United States. This has been primarily the work of the Canadian paleontologist E. Timothy Tozer, who, with the American paleontologist Norman J. Silberling, has provided precisely defined stratotypes for all the recognized North American biozones. The North American zonal scheme, once defined, has since allowed Alpine (western Tethyan) zones to be placed in their proper chronological sequence. This scheme has been adopted here almost in its entirety and is shown in Table 18.

Controversy over boundaries

The exact position of both the Permian-Triassic and Triassic-Jurassic boundaries has been the subject of great controversy for many years. The transition from latest Permian to earliest Triassic is nowhere represented by a continuous (conformable) succession of marine strata containing fossils that are not open to ambiguous age interpretation. The Germanic facies is of little value in the dispute, for here the continental Bunter rests unconformably on Upper Permian Zechstein strata. The marine equivalent of the Bunter in the Alps is the Werfen Limestone that contains the distinctive Lower Triassic bivalve genus *Claraia* near its apparently conformable contact with the underlying Bellerophon Limestone, in which undisputed Permian faunas are found. Recent studies, however, suggest that the lowermost Werfen may contain Permian fossils. In the Himalayas, *Claraia* occurs with the ammonoid *Otoceras* in the so-called *Otoceras* beds, but are these beds Permian or Triassic? A Triassic age is suggested by the presence of *Claraia,* but otoceratids also occur in undisputed Permian strata in the Dzhulfa (Julfa) region in Armenia near the Iran border. It was agreed as long ago as the early 1900s that the Armenian otoceratids were not in the strictest sense identical with *Otoceras* and that the Himalayan *Otoceras* beds should define the basal Triassic. This issue, however, has been raised again by those who regard the *Otoceras* beds as Permian rather than Triassic.

At key localities where apparently conformable sequences occur, such as in Armenia, Pakistan, Kashmir, Arctic Canada, Greenland, Spitsbergen, Tibet, China, Siberia, and northern Alaska, the boundary beds—often of limited thickness—usually contain mixtures of Permian-type and Triassic-type faunas or show evidence of a disconformity or paraconformity (*i.e.*, the strata are parallel but the surface is nearly indistinguishable from a simple bedding plane since no effects of erosion are discernible). It is these transitional beds that are the crux of the boundary problem. In the Salt Range (of Pakistan), for instance, Permian brachiopods are found in close association with undisputed Triassic fossils, suggesting the possibility of Permian relics living in earliest Triassic time. Yet, recent studies suggest that both latest Permian and earliest Triassic strata are missing in this section. In East Greenland mixed faunas occur at the boundary, with Triassic ammonoids in association with Permian productacean brachiopods, but the latter appear to be derived, having been incorporated into Triassic sediments by reworking. A similar situation may prevail at the famous Guryul Ravine section in Kashmir. Studies on new sections in southern Tibet (Selong) and southern China (Mei Shan) have not yet led to agreement on whether there is continuous sedimentation between the Permian and Triassic or a well-disguised unconformity. Tozer supports the latter view and, furthermore, believes that there is evidence of a worldwide unconformity (often well-disguised) at the base of the *Otoceras* zone. He advocates that this level should once again define the Permian-Triassic boundary, since it clearly records a universal geologic event of great significance to marine biotas. Accordingly, he has proposed a stratotype for the boundary at the base of the Blind Fjord Formation of northwestern Axel Heiberg Island in Arctic Canada, where the *Otoceras concavum* zone (equivalent to the *O. woodwardi* zone of the Himalayas) rests disconformably on Permian strata.

The exact position of the boundary between the Triassic and Jurassic has been less contentious but not without its problems. Marine rocks stratigraphically above the Keuper Marl in Germany and the New Red Sandstone in Britain have traditionally been regarded as either uppermost Triassic or lowermost Jurassic. These rocks contain the distinctive bivalve species *Rhaetavicula contorta* but no ammonoids. Rocks of this *R. contorta* zone in northwestern Europe have been correlated with the stratotype of the Rhaetian, the marine Kössen beds in the Rhaetic Alps, mainly on the basis of the common occurrence of *R. contorta.* The Alpine Rhaetian contains a few ammonoids

Rhaetian stratotype

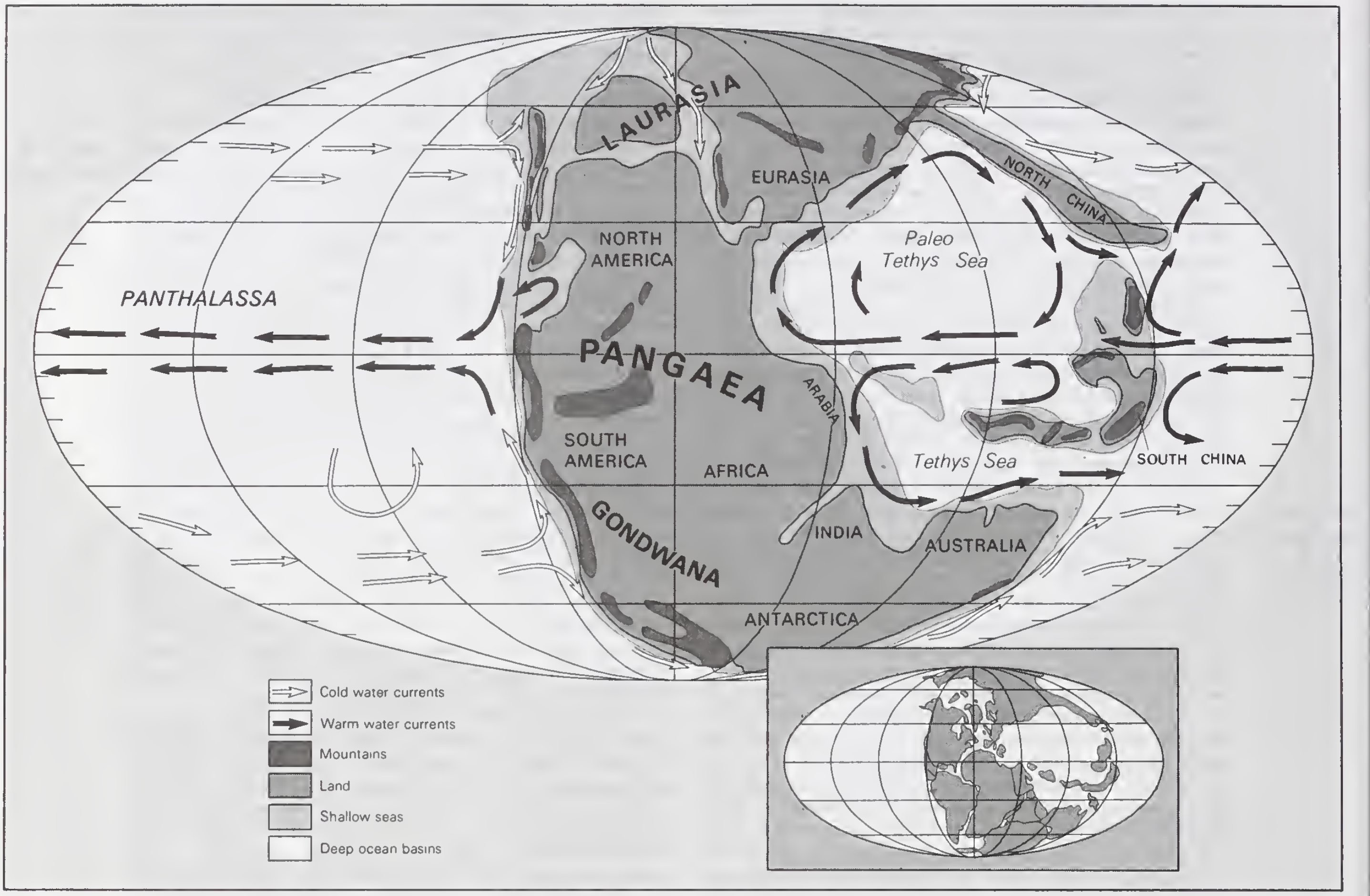


Figure 33: Paleogeography and paleoceanography of Early Triassic time. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset at the lower right.

Adapted from. C.R. Scotese, The University of Texas at Arlington

that are regarded as Late Triassic in affinity but not exclusively Rhaetian. The correlation of the Rhaetian of northwestern Europe with that of the Alps has been questioned, however, and it has been suggested that the former may be lowermost Jurassic in age, preceding the Hettangian *Psiloceras planorbis* zone. While all biostratigraphers would include at least the Alpine Rhaetian in the Triassic, Tozer advocates abandoning the term Rhaetian as a formal stage name and assigning Alpine Rhaetian rocks and their correlatives in North America and elsewhere to the uppermost Norian stage. Since the Rhaetian is so well known, the term has been retained here as a stage name, represented by a single zone, that of ammonite *Choristoceras marshi.*

Triassic environment. *Paleogeography.* As has been noted, at the beginning of the Triassic the present continents of the world were grouped together into one large supercontinent, Pangaea, which covered about one-quarter of the Earth's surface. Pangaea stretched from pole to pole in a narrow belt of about 60° of longitude and consisted of a group of northern continents, Laurasia, and a group of southern continents, Gondwana. The rest of the globe was covered by Panthalassa, the enormous world ocean that stretched from pole to pole and extended to about twice the width of the present-day Pacific Ocean at the equator (Figure 33). Scattered across Panthalassa, within 30° of the Triassic equator, were islands, seamounts, and volcanic archipelagoes, some associated with reef carbonates, which would later be driven into the lithospheric plates on either side of the Panthalassa spreading centre as displaced terranes.

A deep embayment of Panthalassa projected westward between Gondwana and Laurasia along an east–west axis approximately coincident with the present-day Mediterranean Sea. This ancient sea, the Tethys, was later to extend farther westward, as rifting between Laurasia and Gondwana began in Late Triassic time. This seaway would eventually link up with the eastern side of Panthalassa by Middle to Late Jurassic time when Gondwana proceeded to separate from Laurasia. The evidence for these paleogeographic reconstructions comes from many sources, of which paleomagnetic data and the matchups of continental margins, rock types, orogenic events, and distribution of fossil land vertebrates and plants prior to the breakup of Pangaea are the most important. In addition, the recalculated polar-wandering curves for Africa and North America converge between the Carboniferous and Triassic and then begin to diverge in the Late Triassic, indicating the exact time of the onset of separation of these two continents as the Tethys seaway began to open up.

Onset of rifting between Laurasia and Gondwana

Thick sequences of clastic sediments accumulated in marginal geosynclines bordering the present-day circum-Pacific region and the northern and southern margins of the Tethys, while shelf seas occupied parts of the Tethyan, circum-Pacific, and circum-Arctic regions but were otherwise restricted in distribution. Much of the circum-Pacific and the northeastern part of Tethys were active plate margins, as indicated by tectonic and igneous activity, but its northwestern and southern margins were passive during the Triassic.

Significant geologic events. The Triassic Period is characterized by few geologic events of major significance, in contrast to the Jurassic and Cretaceous when Pangaea fragmented and the new Atlantic and Indian oceans opened up. The beginning of continental rifting in the Late Triassic, however, caused stretching of the crust in eastern North America along the Appalachian Mountain belt from the Carolinas to Nova Scotia, resulting in normal faulting in this region. Here grabens received thick clastic sequences, which were later intruded by dikes and sills. In similar fault-controlled basins between Africa and Laurasia evaporites formed in arid or semiarid environments as seawater from the Tethys periodically spilled into these newly formed troughs. Evaporites of Late Triassic and Early Jurassic age in Morocco and off eastern Canada were apparently the result of such tectonism.

Limited orogeny

Mountain building was equally restricted during the Triassic, with relatively minor orogenic activity taking place along the Pacific coastal margin of North America and in China and Japan. The unmetamorphosed nature of the Triassic rocks of the Newark Group indicates that they were formed after the main phase of the Appalachian orogeny in the late Paleozoic.

Paleoclimate. The emergence of the supercontinent Pangaea and the reduction of shelf seas led to widespread aridity over most of the land areas beginning in the Late Permian and continuing throughout the Triassic. The relative paucity of shelf seas has been attributed to the reduced total shoreline of Pangaea and the lack of movement of the continents during this period of relative geologic quiescence. A single large landmass like Pangaea should experience an extreme continental climate, with widespread aridity. Paleoclimatic indicators for such aridity include poorly fossiliferous red sandstones and shales, lithified dune deposits with cross-bedding, salt pseudomorphs in marls, and evaporites, and these are found mainly in low and middle latitudes on the continents. Coal deposits may form in temperate and tropical climates, but their presence invariably indicates humid conditions, with relatively high rainfall responsible for both lush vegetational growth and poor drainage. The resultant large swamps act as depositional basins wherein the decomposing plant material can be gradually transformed into peat. Such humid conditions must have existed in high latitudes during the Triassic, based on the occurrence of coals in Arctic Canada, Russia, Ukraine, China, Japan, South America, South Africa, Australia, and Antarctica.

Homogeneous global climatic conditions

Worldwide climatic conditions during the Triassic seem to have been much more homogeneous than at present. It has been postulated that, because of the large size of Panthalassa, oceanic circulation patterns during the Triassic would have been very simple, consisting of enormous single gyres in each hemisphere (Figure 33). East–west paleotemperature extremes would have been great, with the western margin of Panthalassa much warmer than the eastern side. The permanent westerly equatorial current would have provided warm waters to Tethys, enabling reefs to develop there wherever substrates and depths were favourable. Temperature differences between the equator and the poles would have been less extreme than they are today, the result of which would have been reduced biotic provincialism.

The nature of Triassic fossils and their latitudinal distribution provide additional important evidence of paleoclimates. The biotas of the period are more modern in aspect, and so their life habits and environmental requirements can be reconstructed with relative confidence from comparisons with living relatives. As an example, the presence of colonial scleractinian corals as framework builders in Tethyan reefs of Late Triassic age suggests an environment of low-latitude, warm shelf seas. These seas must have been sufficiently shallow and clear to allow penetration of adequate light for photosynthesis by zooxanthellae (a form of dinoflagellate) inferred to be, perhaps for the first time in geologic history, symbiotically associated with reef-building corals.

Geographic distribution of modern-day animals indicates, with few exceptions, a faunal diversity gradient decreasing poleward in both hemispheres and apparently based on ambient temperature differences between high and low latitudes. Plots of the diversity of ectothermic (cold-blooded) amphibians and reptiles at the present time, for example, show a much higher diversity in low latitudes, reflecting the strong influence of ambient air temperatures on these animals that lack an internal temperature-regulatory system. The evidence from Triassic fossils is equivocal, however. Diversity–latitude plots of Triassic amphibians and reptiles appear to show only a weak diversity gradient, but plots of ammonoids from the upper part of the Lower Triassic show a much stronger gradient. These data suggest that Triassic marine invertebrates were more sensitive to differences in ambient temperature than land vertebrates or that ambient temperature differences were greater in the ocean than on land. There is also the possibility that both these conditions existed.

Triassic life. Periodic large-scale mass extinctions have occurred throughout the history of life; indeed, it is on this basis that the geologic eras were first established. One of the most severe and widespread of such extinction events took place at the end of the Mesozoic; another was at the end of the Paleozoic, and it is the latter that profoundly affected life during the Triassic. A third episode of mass extinctions occurred at the end of the Triassic, drastically reducing some marine and terrestrial groups, such as the ammonoids, mammallike reptiles, and primitive amphibians.

Mass extinctions

The Permo-Triassic extinction was perhaps the most drastic in the history of life on Earth, although it should be noted that many groups were showing evidence of a gradual decline long before the end of the Paleozoic. The trilobites, a group of arthropods long past their zenith, made their last appearance in the Permian, as did the closely related eurypterids. Rugose and tabulate corals became extinct at the end of the Paleozoic. Several superfamilies of Paleozoic brachiopods, such as the productaceans, chonetaceans, spiriferaceans, and richthofeniaceans, also disappeared at the end of the Permian. Fusulinid foraminiferans, useful as late Paleozoic index fossils, did not survive the crisis, nor did the cryptostomate and fenestrate bryozoans that inhabited many Carboniferous and Permian reefs. Gone also were the blastoids, a group of echinoderms that persisted in what is now Indonesia until the end of the Permian, although their decline had begun much earlier in other regions.

Many possible causes have been advanced to account for these extinctions. Cataclysmic events, such as intense volcanic activity and the impact of a celestial body, or more gradual changes brought about by widespread marine regression, oceanic salinity and nutrient fluctuations, climatic cooling, and cosmic radiation, have been proposed to explain the Permo-Triassic crisis. Any theory, however, must take into account that not all groups were affected to the same extent. Whatever the cause, Early Triassic biotas were impoverished, although they progressively increased in diversity and abundance during Middle and Late Triassic times. While the fossils of many Early Triassic life-forms tend to be Paleozoic in aspect, those of the Middle and Late Triassic are decidedly Mesozoic in appearance and clearly the precursors of things to come.

Invertebrates. The difference between Permian and Triassic faunas is most noticeable among the marine invertebrates; the number of families was reduced by half, with an estimated 95 percent of all species disappearing at the Permian–Triassic boundary. The ammonoids were common in the Permian but suffered drastic reduction at the end of that period. Only a few genera belonging to the prolecanitid group survived the crisis, but their descendants, the ceratitids, represented by such Early Triassic genera as *Otoceras* and *Ophiceras,* provided the rootstock for an explosive adaptive radiation in the Middle and Late Triassic. Ceratitids have varying external ornamentation, but all share the distinctive ceratitic internal suture line of rounded saddles and denticulate lobes. The group reached its acme in the Carnian, with more than 150 genera; it declined to less than 100 in the Norian and to less than 10 in the Rhaetian. In the Late Triassic bizarre heteromorphs with loosely coiled body chambers, such as the genus *Choristoceras,* or helically coiled whorls (*e.g., Cochloceras*) evolved. These aberrant forms, however, were short-lived. A small group of smooth-shelled forms with more complex suture lines, the phylloceratids, also arose in the Early Triassic. They are regarded as the earliest true ammonites and gave rise to all post-Triassic ammonites, even though Triassic ammonoids as a whole almost became extinct at the end of the period.

Adaptive radiation of the order Ceratitida

Other marine invertebrate fossils found in Triassic rocks, albeit much reduced in diversity compared with the Permian, include gastropods, bivalves, brachiopods, bryozoans, corals, foraminiferans, and echinoderms. These groups are either poorly represented or absent in Lower

Triassic rocks but increase in importance later in the period. Most are bottom-dwellers (benthos), but the bivalve genera *Claraia, Posidonia, Daonella, Halobia,* and *Monotis,* often used as Triassic guide fossils, were planktonic and may have achieved widespread distribution by being attached to floating seaweed. While the role of colonial scleractinian corals as reef-builders in Middle and Late Triassic structures has already been mentioned, cavities in Rhaetian reefs from Austria were colonized by a cryptofauna of echinoids, foraminiferans, spongiomorphs, and small arthropods. Many successful Paleozoic articulate brachiopod superfamilies became extinct at the end of the Permian, leaving the spiriferaceans, rhynchonellaceans, terebratulaceans, terebratellaceans, thecideaceans, and some other less important groups to continue into the Mesozoic. The brachiopods, however, never again achieved the dominance they held in the benthos of the Paleozoic. Fossil echinoderms are represented in the Triassic by crinoid columnals and the echinoid *Miocidaris,* a holdover from the Permian. The crinoids had begun to decline long before the end of the Permian, by which time they were almost entirely decimated, with both the flexible and camerate varieties dying out. The inadunates survived the crisis; they did not become extinct until the end of the Triassic and gave rise to the articulates that still exist today. Coccolithophores, an important group of living marine pelagic algae, made their first appearance during the Late Triassic, while dinoflagellates underwent rapid diversification during the Late Triassic and Early Jurassic.

Vertebrates. Vertebrate animals appear to have been less affected by the Permo-Triassic crisis than were invertebrates. The fishes show some decline in diversity and abundance at the end of the Paleozoic, with acanthodians becoming extinct and elasmobranchs much reduced in diversity. Actinopterygians, however, continued to flourish during the Triassic, gradually moving from freshwater to marine environments, inhabited by subholostean ray-finned fishes, which were intermediate between palaeoniscoids and holosteans. The shellfish-eating hybodont sharks, already diversified by the end of the Permian, continued into the Triassic. The fossils of marine reptiles, such as the shell-crushing placodonts (which superficially resembled turtles) and the fish-eating nothosaurs, occur in the Muschelkalk of the Germanic facies. The latter forms, sauropterygians, did not survive the Triassic, but are ancestral to the large predatory plesiosaurs of the Jurassic. The largest inhabitants of Triassic seas were the early ichthyosaurs, superficially like dolphins in profile and streamlined for rapid swimming. These efficient hunters were equipped with powerful fins, a long-toothed jaw, and large eyes and may have preyed upon some of the early squidlike cephalopods known as belemnites. There also is evidence that these unusual reptiles gave birth to live young.

Giant marine reptiles

On land, the vertebrates are represented in the Triassic by labyrinthodont amphibians and reptiles, the latter consisting of cotylosaurs, therapsids, eosuchians, thecodonts, and protorosaurs. All these tetrapod groups suffered a sharp reduction in diversity at the close of the Permian; 75 percent of the early amphibian families and 80 percent of the early reptilian families disappeared at or near the Permian–Triassic boundary. While early Triassic forms were still Paleozoic in aspect, new forms appeared throughout the period, and by late Triassic times the tetrapod fauna was distinctly Mesozoic in aspect. The mammallike reptiles, or therapsids, suffered pulses of extinctions in the Late Permian. The group survived the boundary crisis but became extinct by the end of the Triassic, possibly due to competition from more efficient predators, such as the thecodonts. The first true mammals appeared in the Late Triassic. Although their fossilized remains have been collected from the Rhaetian bone bed in Britain, the evolutionary transition from therapsid reptiles to mammals at the close of the Triassic is nowhere clearly demonstrated by well-preserved fossils. Other modern groups with ancestral forms appearing for the first time in the Middle and Late Triassic include lizards, turtles, rhynchocephalians, and crocodilians.

Emergence of the first true mammals

First encountered in the Early Triassic, the thecodonts became common during the Middle Triassic, but disappeared before the beginning of the Jurassic. Typical of this group of archosaurs (or ruling reptiles) in the Triassic were small bipedal forms belonging to the pseudosuchians. These were swift-running predators that had erect limbs directly under the body, which made them more mobile and agile. This group presumably gave rise to primitive dinosaurs belonging to the saurischian and ornithischian orders during the Late Triassic to Early Jurassic. The early dinosaurs were bipedal, swift-moving, and relatively small compared to later Mesozoic forms, but still reached lengths of more than six metres. The group was to achieve much greater importance later in the Mesozoic, resulting in the era being informally called the "Age of Reptiles."

Some of the earliest lizards may have been the first vertebrates to take to the air. Gliding lizards, such as the small Late Triassic *Icarosaurus,* are thought to have developed an airfoil from skin stretched between extended ribs, allowing short glides similar to those made by present-day flying squirrels. These forms became extinct at the end of the Triassic, their role as fliers being taken over by the pterosaurs (flying reptiles) of the Jurassic and Cretaceous.

Plants. Land plants were relatively unaffected by the Permo-Triassic crisis. The dominant understory plants in the Triassic were the ferns, while most middle-story plants were gymnosperms (those having exposed seeds) that belonged to the still-extant cycadeoids and Ginkgoales. The upper story of Triassic forests consisted of conifers; their best-known fossil remains are preserved in the Upper Triassic Chinle Formation in the Petrified Forest National Park of northeastern Arizona. While extensive forests did exist during the Triassic, widespread aridity on the continents limited their areal extent, resulting in generally poor development of fossil floras during this period. In the southern continents the Permian *Glossopteris–Gangamopteris* seed-fern flora was replaced by a Triassic one dominated by *Dicroidium.* This pteridosperm genus was part of an extensive Gondwanan paleoflora that was discovered in the Triassic Molteno Formation of southern Africa and elsewhere. (This paleoflora extended from 30° S to well below 60° S.) In the Northern Hemisphere a subtropical-to-warm temperate Eurasian (Euramerican) flora lay in a belt between about 15° N and 55° N, while to the north was the temperate Siberian (Angaran) flora, extending to within 10° of the Triassic North Pole. Few fossil remains exist from the Triassic for the equatorial zone between 15° N and 30° S, however. (A.Lo.)

JURASSIC PERIOD

General considerations. The Jurassic is the second of three periods of the Mesozoic Era. On the basis of radiometric measurements, it extended from about 208 to 144 million years ago (see Table 4). The rocks that originated during this interval of time compose the Jurassic System. In many regions of the world, the Jurassic lies directly over the Triassic strata and, in turn, is often overlain by those of the Cretaceous. The Jurassic was named early in the 19th century by the French geologist and mineralogist Alexandre Brongniart for the carbonate terrane of the Jura Mountains between France and Switzerland. The presence of similar fossils was used to establish a correlation between the Jura carbonates and the oolite limestones of England.

Major subdivisions. The Jurassic is generally subdivided into three parts—Early, Middle, and Late—after the chronology of the German geologist Leopold Buch. Later, F.A. Quenstadt and C.A. Oppel used the names Lias, Dogger, and Malm, respectively, for these units, which were also referred to as the Black, Brown, and White Jurassic sequences (see Table 19).

Jurassic stages

Each of the triad was further subdivided into stages by Oppel in Germany and Brongniart, A.D. d'Orbigny, and other geologists elsewhere in western Europe, who named the stages after existing rock or geographic names. For example, the Hettangian Stage was named by the Swiss geologist Eugène Renevier after Hettange-Grande of the Lorraine in eastern France (Table 19). Eventually a biostratigraphic basis for recognizing the stages was established through the determination of zone fossils, index fossils, or fossil assemblages, and in some cases fossils accompanying distinctive rock types (Table 19). In this way, d'Orbigny named the following stages: Sinemurian (after Sémur in the south of France), Toarcian (Thouars in western France), Bajocian (Bayeux in Normandy), Bathonian (Bath, Eng.), Callovian (Kellaways, Eng.), and Kimmeridgian (Kimmeridge, Eng.). Brongniart named the Oxfordian and Portlandian after Oxford and Portland Isle in southern England, and Oppel derived Pliensbachian from that location in Germany.

Correlation of fossils and rock layers made it possible for William Smith to prepare a geologic map that showed the aerial distribution of Mesozoic and Cenozoic rocks in southern England, and Georges Cuvier accomplished the same task in northwestern France. These areas are now commonly referred to as the London and Paris basins, respectively. The Jurassic underlies younger rocks in the interior of both basins. The Tithonian Stage was added by Oppel for the uppermost Jurassic of the Alpine region, and Sergey Nikolaevich Nikitin named the Volgian Stage for the upper Jurassic in the Volga basin in Russia. The underlying Triassic and overlying Cretaceous strata are distinguished from those of the Jurassic on the basis of fossils and unconformities.

Development of geosynclines

Distinctive features. Jurassic rocks are widely distributed as part of both oceanic and continental crust. They include deep-sea sediments, coral limestone, carbonate platforms, volcanic island arcs, and interarc basin deposits, as well as terrestrial sediments and extensive plutonic intrusions. While lithospheric plate divergence and continental rifting mark the late Triassic before 208 million years ago, breakup of the supercontinent Pangaea and the opening of the ocean basins date from Early Jurassic time. Geosynclines developed along the new continental margins. Some of these, such as the miogeoclines (prograding wedges of shallow-water sediments) on the trailing, or passive, margin of eastern North America and the Gulf Coast, persist today. Others, like the Tethyan geosyncline of southern Eurasia, its counterpart along the northern margin of Gondwana, and the West Coast geosynclinal complex of western North America—all of which occupied the leading edges of the continents—have been converted to mountain ranges through collisions between continental plates or between oceanic and continental plates.

Jurassic plate tectonics accounts for the accretion to continental plates of 5,000 to 16,000 metres of sediments from continental margins, oceanic crust, island arcs, and arc-related basins, as well as from igneous intrusions. In North America, the Nevadan orogeny of the Late Jurassic involved the collision of an island-arc complex with the western margin of the continent. Because of plate collisions (and the attendant subduction of oceanic crust) and the opening of new ocean basins, there is no ocean floor in the present-day basins that is older than Jurassic age. Due to late- and post-Mesozoic plate collisions between Africa and Europe and the fragments of Gondwana and Asia, Jurassic carbonates of the Tethys Sea lie in extensive thrust sheets, called nappes, which blanket the crystalline core of the region in the mountainous uplands from the Alps to the Himalayas.

Sedimentation continued during the Jurassic in the Mesozoic rift basins of continental eastern North America (see Figure 35), southwestern Europe, and southern Africa—basins that are similar to the expanding rift basins of present-day East Africa. These deposits include red beds and lacustrine shales interbedded with mafic lavas (*i.e.*, pillow lavas where extrusion was into standing water) and coarse-grained alluvial fan conglomerates. Epicontinental seas left sandy coastal plains, shallow-water marine beds, and evaporites with fluvial and deltaic facies intermixed.

Adapted from C R Scotese, The University of Texas at Arlington

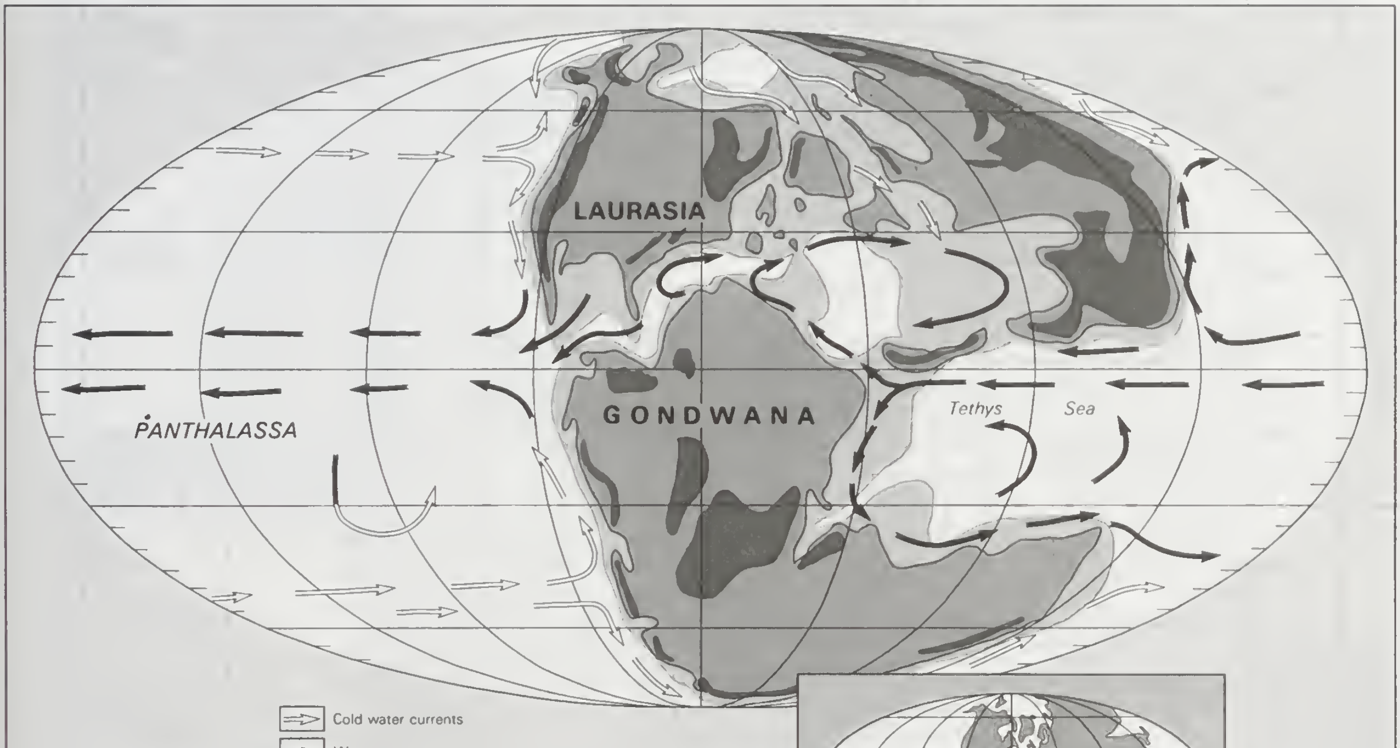


Figure 34: Distribution of landmasses, mountainous regions, shallow seas, and deep ocean basins during the Late Jurassic. Included in the paleographic reconstruction are cold and warm ocean currents. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset.

Table 19: Selected Stratigraphic Columns with Correlations and Zonations for Europe and North America

series	European stages (*Arkell, 1946*)	English formations (*Arkell, 1933*)	Northwest Europe standard zones (*Arkell, 1946*)	characteristic fossils in the Western interior region of the United States	Goodsprings Quadrangle and Muddy Mountains, Nevada	Lees Ferry, Arizona
overlying beds					Tertiary Upper Cretaceous	
	Tithonian/Volgian	Purbeck beds		?		
Upper Jurassic	Portlandian	Portland beds	*Titanites giganteus*			
			Kerberites okusensis			
			Glaucolithites gorei	*Vetulonaia* species and *Gyraulus veternus*		
			Zaraiskites albani			
	Kimmeridgian	Kimmeridge clay	*Pavlovia pallasioides*			
			Pavlovia rotunda			
			Pectinatites pectinatus			
			Subplanites wheatleyensis			
			Subplanites species			
			Gravesia gigas			
			Gravesia gravesiana			
			Aulacce tephonus pseudomutabilis			
			Rasenia mutabilis			
			Rasenia cymodoce			
			Pictonia baylei			
	Oxfordian	Corallian beds	*Ringsteadia pseudocordata*	?		
			Decipia decipiens			
			Perisphinctes cautisnigrae			
			Perisphinctes plicatilis			
		Oxford clay	*Cardioceras cordatum*	*Cardioceras* species		
			Quenstedtoceras mariae	*Cardioceras cordiforme*		
	Callovian		*Quenstedtoceras lamberti*	*Quenstedtoceras collieri*		?
			Peltoceras athleta	?		Entrada sandstone
			Erymnoceras coronatum			
			Kosmoceras jason			
		Kellaways beds	*Sigaloceras calloviense*	*Kepplerites mclearni* *Kepplerites cf. tychonis*		?
			Proplanulites koenigi	*Gowericeras subitum*		Carmel formation
		Cornbrash beds (limestone)	*Macrocephalites macrocephalus*	*Arcticoceras*		
Middle Jurassic	Bathonian		*Clydoniceras discus*	*Arctocephalites*		
		Great oolite (limestone)	(not yet determined)			
	Bajocian		*Parkinsonia parkinsoni*	*Stemmatoceras* and *Chondroceras* (not zoned)		
		Inferior oolite (limestone)	*Stephonocares humphr lesianum*			
			Otoites sauzei			
			Sonninia sowerbyi	?		?
			Ludwigia murchisonae			
			Lioceras opalinum			
Lower Jurassic	Toarcian	Upper Lias (shale, sandstone)	*Lytoceras jurense*		?	?
			Hildoceras bifrons		Navajo sandstone "Aztec sandstone"	Navajo sandstone
			Harpoceros serpentinum			
			Dactylioceras tenuicostatum			
	Pliensbachian	Middle Lias (shale)	*Poltopleuroceras spinatum*			
			Amaltheus margaritatus			
		Lower Lias (shale)	*Prodactylioceras davoei*			
			Tragophylloceras ibex			
			Uptonia jamesoni			
	Sinemurian		*Echioceras raricostatum*			
			Oxynoticeras oxynotum			
			Asteroceras obtusum			
			Arietites turneri			
			Arnioceras semicostatum			
			Coroniceras bucklandi			
	Hettangian		*Scamnoceras angulatum*		?	?
			Psiloceras planorbis			

Table 19: (continued)

Source: Derived in part from R.W. Imlay, *Correlation of the Jurassic Formations of North America, Exclusive of Canada* (1952), originally published in the *Geological Society of America Bulletin*, 63:953–992.

Kayenta, and Rough Rock, Arizona	Beciabito Dome, New Mexico and Red Rock, Arizona	Gulf Coast	England	Russian platform	Northwest Germany	Southern Germany (Swabia)	Jura Mountains	Southern France (Ardèche)
Cretaceous	Cretaceous							
			A Lower Purbeck Beds and Lower Spilsby Sands	A Upper Volgian	Lower Serpulite		A Upper Tithonian Limestones	A Upper Tithonian Limestones (Ardèche)
1 Morrison formation	1 Morrison formation	Cotton Valley Group	A Portland Beds	greensand, marl ⊖	Münder Marls		beds of Purbeck Facies	
			A Upper Kimmeridge Clay	A Lower Volgian		Neuburg Beds ζ	limestones of Mont Salève, etc.	Middle and Lower Tithonian
			A Middle and Lower Kimmeridge Clay	A clays	A Gigas Beds Limestone 〰	Solnhofn–Plattenkalke	A Lower Tithonian Limestones	A Limestones
?	?				A clays Corallian Oolite		A limestones with coral reefs	A limestones of Crussol
Curtis or Wanakah formation: ?	Curtis or Wanakah formation: Sandstone and shale	Smackover Formation	A Corallian Beds	A clays	A Heersum Clays	A A Malm or White Jura ε δ γ β α	marls	A marly limestones
? Todilto limestone member ?	? Todilto limestone member ?	Eagle Mills Formation	A Oxford Clay		A clays	A ζ	A oolite ironstone and limestone ⊖	A marls and dolomite
3 Entrada sandstone	3 Entrada sandstone			A clays	A succession near Hildesheim			
? 4 Carmel formation	? 4 Carmel formation		A Kellaways Beds					
			A Cornbrash Limestones ⊖		Fuller's earth	Dogger or Brown Jura ε		A limestones ⊖
			A Great Oolite Limestones, etc.	sands	A Bielefeld Succession A clay	A δ γ β α	A oolite limestone and A marls	A
			A Inferior Oolite Limestones	Bajocian and Toarcian in Donets Basin	succession near Hanover		A clays and oolite ironstone	
?	?				A	A ζ ε		A ferruginous oolite ⊖
? 5 Navajo sandstone	? 5 Navajo sandstone (or Negget sandstone)		A Upper Lias Shales, Sands			γ δ		
			A Middle Lias Shales, Ironstone		A	A Lias or		A
			A A Lower Lias Shales		Lias A	A Black Jura β α	A Middle and Lower Lias Black Marls	
? Kayenta Formation	? Kayenta Formation / absent at Red Rock		A		A	A	A	
? Wingate sandstone	? Wingate sandstone							

Key: A ammonite correlations, ⊖ thin but significant deposits, ζ through α, Quenstedt's classical subdivisions of Swabian Jurassic; 〰 strong folding.

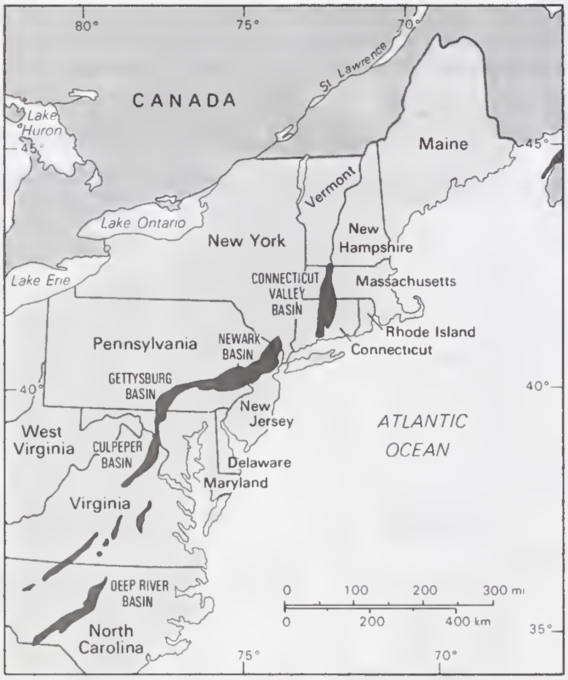


Figure 35: Distribution of Mesozoic rift basins along the northern Atlantic coast of the United States.

Adapted from *Earth History and Plate Tectonics* by Carl K. Seyfert and Leslie A. Sirkin. Copyright © 1979 by Carl K. Seyfert and Leslie A. Sirkin. Reprinted by permission of Harper Collins Publishers Inc.

Deposition on oceanic crust began in marine basins where evaporation prevailed during Middle Jurassic time (marine sediments of Early Jurassic age are entirely absent on oceanic crust).

Coal and petroleum deposits

Economic significance. Epicontinental seas invaded continental interiors to varying degrees between the Early and Late Jurassic, leaving extensive beds of clay, limestone, sandstone, ironstone, and coal. These deposits have provided economically important resources in many continental interior regions. For example, clay and limestone have been used for brick, cement, and other building materials in various areas of continental Europe. Iron ore and coal for steelmaking are prevalent in western Europe and England. Coal deposits also are found in Siberia and Asia, as, for example, in the Late and Middle Jurassic Yenan Formation in the Ordos Basin of China. Petroleum is produced from strata trapped against salt domes of Jurassic age in the Gulf Coast province of the United States and from the marginal basins of western North America. Oil also is found in northern Germany and Russia, and salt is mined in both the United States and Germany. Lithographic limestone is quarried in Germany as well. Jurassic igneous rocks have yielded chromite in eastern Europe and uranium and gold concentrations in placer deposits on the slopes of the Sierra Nevada in the western United States.

Characteristic fauna and flora. Distinctive Jurassic lifeforms range from microscopic unicellular organisms known as coccolithophores (or coccoliths) to the most diverse animal group, the dinosaurs. After the Late Triassic to Early Jurassic extinctions left only a few ammonoid cephalopod and reptile genera, the ammonoids rebounded so successfully that their remains have come to serve as marine guide fossils in the biostratigraphic zonation of Jurassic stages. Dinosaurs and other reptiles emerged to dominate the land, sea, and sky. The first birds and new varieties of gymnosperms, along with extensive reef-building and other invertebrate faunas, provided Jurassic life with added complexity.

Breakup of Pangaea

Jurassic rocks. *Occurrence and distribution.* Pangaea began to break up as North America separated from Eurasia and Gondwana, with Africa subsequently splitting off from India, Australia, and Antarctica. As a result of this intense tectonic activity, alkalic igneous rocks were emplaced in the New England region of North America, and mafic volcanics were extruded in southern Africa during the Early Jurassic (Sinemurian). By Late Jurassic time, Africa had separated from South America and Australia and Antarctica had broken away from India. At the same time, the Iberian Peninsula rotated away from Europe, while Alaska rotated away from Canada.

Evidence of these plate movements comes from radiometrically dated igneous intrusions, oceanic sediments (the oldest sediments on the oceanic crust in the Atlantic Basin are of Callovian age—about 169 to 163 million years old—and the oldest geosynclinal sediments between North America and Africa are approximately 182 million years old), and magnetic anomalies, the oldest of which in the Atlantic is dated at about 153 million years. In the North Atlantic, the separation of North America and Greenland began in Late Jurassic (Kimmeridgian) time, as shown by the emplacement of dikes between 162 and 138 million years ago. The Late Jurassic continental breakup is associated with tectonic activity in North America during the Nevadan orogeny. The fact that South America separated from Africa between 150 and 130 million years ago has been determined by age measurements of Late Jurassic mafic volcanics, magnetic anomalies, geosynclinal sediments, and alkalic igneous intrusions. India began separating from Australia and Antarctica during the Kimmeridgian.

Eastern North America

In eastern North America, Late Triassic–Early Jurassic extensional basins became filled with synrift deposits (red beds and others produced during continental rifting), and pillow lavas were extruded into lake basins. The Upper Newark Supergroup of these basins is Early Jurassic in age, based on potassium–argon ages of 185- to 194-million-year-old basaltic flows like the Watchung Flows of the Newark Basin. More than 150 metres of Lower Jurassic cyclic fluvial and lacustrine beds were deposited in the Culpeper Basin in northern Virginia. Middle Jurassic volcanoclastic rocks have been found beneath continental shelf sediments on the New England margin of North America. Upper Jurassic marine sediments include clastics interfingering with carbonates in the Atlantic and Gulf Coast miogeoclines that formed on the initial continental margins of North America after its separation from Gondwana. Middle Jurassic strata include evaporites (*e.g.*, Louann Salt), red beds, carbonates, and shelf-margin reefs. The Smackover Formation of the Gulf Coast sequences is a geosynclinal sedimentary unit typical of this interval. Plate activity continued with seafloor spreading and the development of the Caribbean microplate. Paleomagnetic studies of Early Jurassic rocks in eastern North America indicate normal polarity and a correlation with similar magnetostratigraphy in Africa and present-day Russia.

Western continental interior of North America

An Early Jurassic marine transgression, the so-called Sundance sea, covered the western continental interior of North America, depositing thin sandstone, shale, and limestone beds from the Arctic almost to the Gulf of Mexico with accumulations of generally a few hundred metres. Distinctive stratigraphic units include the Early Jurassic Navajo Sandstone, a variable coastal plain sequence that incorporates cross-bedded, dune-formed sandstone and red beds, along with the Late Jurassic Morrison Formation, which is a continental clastic wedge of lacustrine and fluvial mudstone, siltstone, sandstone, and conglomerate that was built to the northeast over the deposits of the epicontinental sea. Uplift of the continental interior occurred between central Arizona and southern California during the interval from the Late Triassic to the Middle Jurassic. Dinosaur fossils (including some tracks) and associated plant and invertebrate fossil remains are especially abundant in the Morrison Formation. Deposition on the Arctic margin of North America was principally of fine-grained marine clastics and a thin continental clastic wedge of sandstone and conglomerate. Early Jurassic deformation was primarily compressional, and deposition continued through Portlandian time—that is to say, into the Late Jurassic.

On the western continental margin of North America, volcanic island arc–continent collisions took place along what is now the foothills of the Sierra Nevada on the eastern edge of California during the Nevadan orogeny. Deformation of the Foothills Terrane in the Sierras occurred from 160 to 150 million years ago. Jurassic ophiolites are dated between 200 and 163 million years, as are intrusive plutons (batholiths and various other large igneous bodies) like the granodiorites of the western Sierras. Plutons in the central Sierras are dated at 150 million years. Volcanics, including pillow lavas, also were emplaced at this time.

During the Middle Jurassic, accretion of oceanic crust in the Coast Range geosyncline involved a basaltic-to-ultramafic arc terrane and ophiolites above an east-dipping subduction zone in the Klamath Mountains of the Oregon–California boundary. This formed Upper Jurassic slate, graywacke, and chert, as exemplified by the Galice Formation and ophiolites of the western Klamath Mountains in southwestern Oregon and the Mariposa Formation in northern California. The Franciscan Formation of the Coast Ranges represents an accretionary wedge of sediments that accumulated in trenches in the plate-collision zone. It is composed largely of deep-water marine deposits called flysch, graywacke, banded radiolarian chert, volcanoclastics, mélange (a mixture of rock fragments of different type, character, and origin), and low-grade metasediments and metavolcanics (metamorphosed sediments and metamorphosed igneous rocks, respectively), together with pillow basalts and greenstone. An Early Jurassic island arc existed in southern Alaska, and a Middle Jurassic back-arc basin and sedimentary mélange—namely, the Coloradito Formation—is found along the western coast of Baja California between the Vizcaíno Peninsula and Cedros Island. The Eugenia Formation of the Late Jurassic represents a younger back-arc basin sequence in the same region. Mafic dikes in northern California also may very well indicate a Late Jurassic volcanic arc. Oceanic sediments from the flat expanse of the deep seafloor known as the abyssal plain and from the continental margin were mixed with carbonates from Pacific atolls to form the accretionary molasse (*i.e.*, thick shore deposits) of the Late Jurassic.

East of the collision zone, deposition of sediments in fore-arc basins of the continental margin gave rise to the Great Valley Group of fine-grained sedimentary rocks. The western margin of the continent evolved into a more extensive system of fore-arc and interarc basin structures of late Kimmeridgian time with an accumulation of up to 16,000 metres of sediment. Subduction of a microplate, known as the Farallon Plate, resulted in Late Jurassic accretion of oceanic crust, arc–arc and arc–continent collisions, emplacement of granitic intrusions, and east-directed thrust faults from northern California to British Columbia due to compressional movement related to the Nevadan orogeny. This process of microplate accretion has incorporated more than 50 Jurassic exotic terranes (see above) to the western Cordilleran orogenic belt. These exotic terranes, which are typified by the Cache Creek Terrane and the Wrangellia Terrane of the Middle Jurassic, consist of segments of oceanic crust that may have originated in tropical regions of the Pacific and were accreted to western North America between southern Alaska and eastern Oregon.

Eurasia and Gondwana

The Tethyan geosyncline on the margins of both Eurasia and Gondwana accumulated thick sequences of Jurassic rock ranging from carbonates and evaporites to sandstone. The carbonates include shallow-water marl, limestone and reefs, and deep-water siliceous limestone. The presence of radiolarian cherts strongly indicates a deep-water origin for some of the sediments. Breccia, found today in nappes of the Alps, was thrust up from submarine rifts, and flysch deposits show the presence of rapidly sinking orogenic basins.

The magnetostratigraphy of Late Jurassic limestone in southern Europe shows normal polarity in the upper part of the M18 magnetic anomaly and reversed polarity in anomaly M17 in the vicinity of the Jurassic–Cretaceous boundary (see Table 4). Based on the first appearance of the nannoplankton fossil *Nannoconus colomi*, the Jurassic–Cretaceous boundary falls at the M17–M18 boundary. Alternatively, the boundary based on the ciliophoride *Calpionella alpina* correlates somewhat differently with magnetic anomalies and is placed at the base of the M18 normal polarity zone.

In the continental interior of Eurasia, shallow seas deposited a complex of carbonates: the limestones, reefs, and marls of the Jura Mountains and of southern France and England, and the fossiliferous, fine-grained lithographic limestone of Germany. Clastic facies include the Lias (Early Jurassic) shales of western Europe, the Oxfordian clays of England and Germany, and the clays of the Russian Platform. While the Arctic region was primarily a clastic province dominated by clay-rich rocks, shale, siltstone, sandstone, and conglomerate, the geosynclinal Pacific margin of Asia developed volcanic island arcs and associated basins from Japan (where thick oceanic crust accreted to the arc) all the way to Indonesia. Both the Australian and Siberian continental interiors received thick continental deposits that include fluvial, eolian, and lacustrine beds.

The rifting and diverging segments of Gondwana variously collided with the adjacent Pacific crust, while rifts in the continental interiors poured out vast amounts of flood basalts—namely, the Ferrar Dolerite of Antarctica, the São Bento Dolerite of South America, the Rajmahal Dolerite of India, and a correlative dolerite in South Africa. Late Jurassic volcanics and deformation extended to New Zealand, as Australia and New Zealand collided with the Pacific Plate. Radiolarian stratigraphy indicates that the Esk Terrane is a tectonic mélange with limestone blocks that was accreted to New Zealand between Late Triassic and Late Jurassic time. In the southern Tethyan region, fine-grained sediments accumulated to form thick shale deposits. Arc–continent collisions in the Andean geosyncline during the Late Jurassic deformed graywackes and volcanics, and volcanism extended into the continental interior of South America. Africa and Madagascar began to separate, eventually opening an oceanic trough for deposition of marine sediments.

Ocean basins

The oldest oceanic sediments date from the mid-Jurassic. Seafloor spreading (and magnetic anomalies) began about 147 million years ago. The Indian Ocean also began to open at this time as India separated from Australia and Antarctica. The oldest crust of the Pacific Basin dates from the Late Jurassic.

Jurassic environment. *Paleogeography.* The continents were still grouped closely in the Pangaean configuration during the Early Jurassic, although divergence in the Atlantic axis was under way at this time. The formation of Pangaea had caused geosynclines to develop on all the coastal margins of the constituent continents. Passive continental margins existed in the Arctic, with a clastic depositional regime, along eastern North America and western Europe, both of which acquired miogeoclinal basins as separation took place, and along the Tethyan geosyncline between Eurasia and Gondwana, which, with its medial deep-sea component, persisted through this interval. Active plate collisions took place around the Pacific rim: the western margins of the North American and South American plates collided with the eastern edge of the Pacific Plate and its microplates and exotic terranes, while Asia, from Japan to Indonesia, and New Zealand collided against the western margin of the Pacific Plate.

Neither of the Jurassic poles was occupied by a landmass. The South Pole lay south of Antarctica, only part of which extended south of 60° S. Most of Antarctica and Australia, as well as the southern tip of South America and Africa, were north of 60° S. The North Pole lay engulfed in the Arctic Sea, and only northernmost North America, along with Greenland and Siberia, extended north of 60° N. The paleoequator cut through the Pacific Ocean and the Tethys Sea, dissecting Gondwana and slicing across the northern edge of South America. As a result of this geographic distribution, tropical and temperate conditions (*i.e.*, those typical of areas between 60° N and 60° S) prevailed on most of the landmasses.

Reefs existed between 30° N and 30° S; most red-bed, evaporite, and dune deposits also occurred in this region. Coal deposits were extensive in northwestern North America, southwestern Europe, much of Siberia, Asia (including China), and Australia. Many of the continental deposits resulted from fluctuations along the margins of the epicontinental seas during the Middle and Late Jurassic, as did the development of multiple interfingering transitional and continental sediments, which in some areas formed cyclothems.

Jurassic spreading centres and mid-oceanic rifts formed between North America and Eurasia, North America and Gondwana, and Eurasia and Gondwana. Such centres and rifts also developed between the various segments of Gondwana itself.

Paleoclimate. Jurassic climates can be deduced from paleogeographic reconstructions that reveal the location of reefs, red beds, evaporites, and dune sandstone. The first three are all indicative of tropical to subtropical conditions, while the presence of the latter suggests semiarid to arid regions and generally subtropical climate. It has been determined that widespread sand sheets, or ergs, covered the desert plains adjacent to the epicontinental seas of the mid-latitude continental interior of North America from Utah to Arizona. Such vast accumulations of sand gave rise to thick eolian sandstone typified by the Early Jurassic Wingate Sandstone. Extensive salt deposits are associated with regions of great aridity, such as the sabkhah (also spelled sebkha) areas that are found near the Red Sea today.

Plate movements

The presence of these indicator rock types in modern temperate, boreal, or polar latitudes indicates that the continental plates moved away from the paleoequator after their deposition during the Jurassic. There is no evidence of glaciation or polar ice caps in the Jurassic, perhaps owing to the lack of a continental landmass in a polar position (see Figure 34). Warm, sun-lit oceans, including the epicontinental seas, abounded in marine life, especially in reef-building invertebrates and the extensive reef infaunas. The widespread development of terrestrial forests and the thick deposits of coal show that most of the existing landmasses were in temperate to tropical climatic regimes with adequate moisture to support the development of plant life.

Jurassic life. *Protists and invertebrates.* Among the prominent marine life-forms of the Jurassic were the Protista, including the foraminiferans, and radiolarians. Foraminiferans, particularly the large-sized varieties in the family Lituolidae, were abundant. Benthic protozoans, planktonic foraminiferans (including the first globigerinids), and radiolarians formed deep-ocean oozes and have provided evidence of the latitude of origin and age of some exotic terranes. Other abundant protists were the flagellates, coccolithophores (calcareous platelet-forming organisms that appeared in the Early Jurassic), dinoflagellates (a Middle Jurassic arrival), hystrichosphaerids (protists of uncertain origin but linked to the dinoflagellates), and the ciliophorans. The latter include the tintinnids, an important group of limestone-secreting organisms of the Late Jurassic particularly in the Tethyan region, and the calcareous calpionellids, important zone fossils, especially for the Upper Jurassic.

Reefs of the Tethys Sea were formed of siliceous sponges and stromatolites, as well as of corals, and had an extensive infauna (see Figure 36). Sponge spicules, along with stromatolites, are prevalent in some limestones in the continental interior of Europe during the Early Jurassic. *Corneyella,* a genus of thick-walled calcareous sponge, emerged in the Upper Jurassic of Germany, and *Pachyterchisma,* a genus of siliceous sponge, also appeared in the Upper Jurassic. Coral reefs were widespread throughout the Tethyan region, with the Anthozoa, mainly the order Scleractinia, dominant during the Late Jurassic in Europe and Africa. A number of scleractinian corals, such as the genera *Actinarea* and *Comophyllia,* were named by d'Orbigny. Scyphozoans, the medusae or jellyfish, are represented in the inland seas of Europe. Genera include *Medusina* and *Rhizostomites.* Large scyphozoan jellyfish are common in Upper Tethyan reef sequences. The contorted, calcareous tubes of serpulid worms are often found clustered in reef-forming masses in the Jurassic carbonates of northwestern Europe.

Few brachiopod families, such as the spiriferids, survived the Paleozoic. Nevertheless, two groups, the rhynchonellids and the terebratulids, evolved in the shallow marine environments and survive today. The spiriferids, on the other hand, became extinct during the Jurassic. Rhynchonellids were widespread in the Jurassic of Europe and North America. Some genera are *Homoeorhynchia* from the Lower to Middle Jurassic of the Alps and western

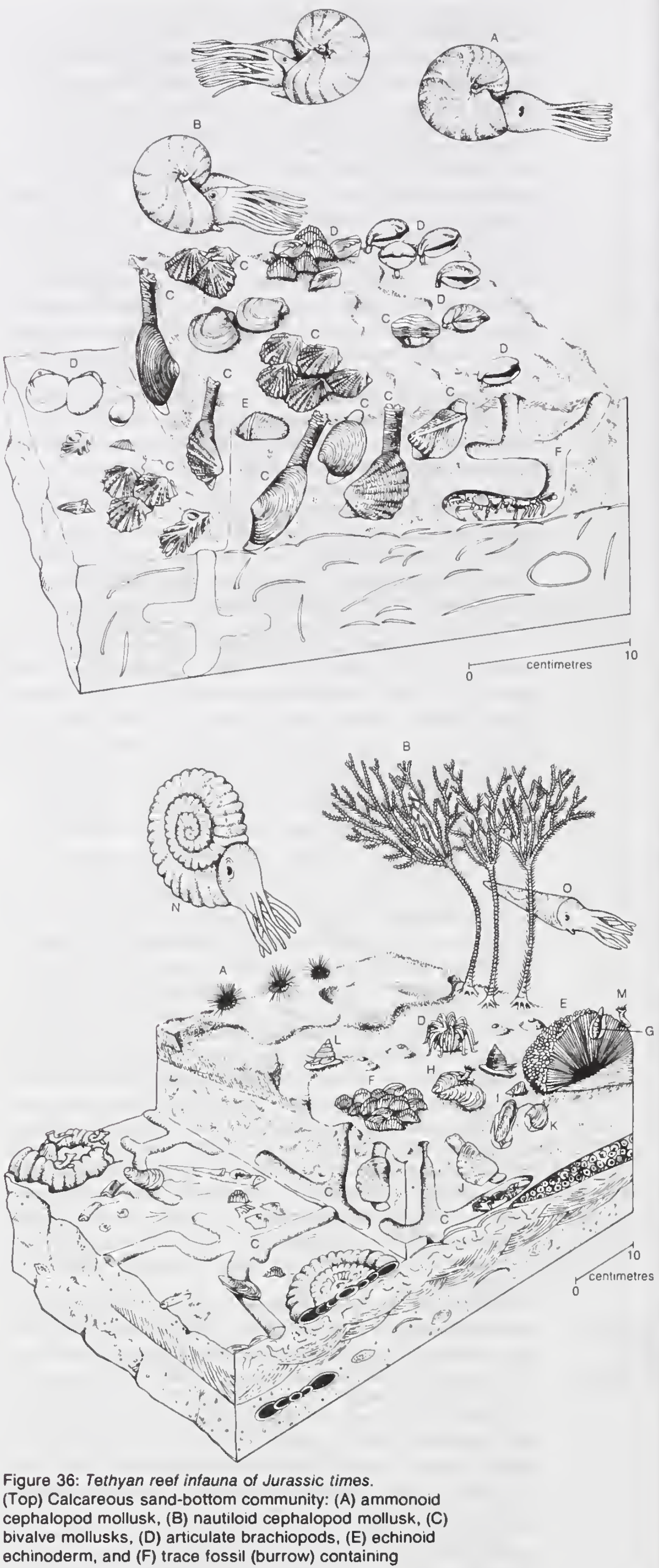


Figure 36: *Tethyan reef infauna of Jurassic times.*
(Top) Calcareous sand-bottom community: (A) ammonoid cephalopod mollusk, (B) nautiloid cephalopod mollusk, (C) bivalve mollusks, (D) articulate brachiopods, (E) echinoid echinoderm, and (F) trace fossil (burrow) containing crustacean. (Bottom) In-life restoration of calcarenite invertebrate community: (A) echinozoan, (B) crinoid, (C) crustacean feeding/dwelling traces, (D–E) scleractinian corals, (F) articulate brachiopod, (G–K) bivalve mollusks, (L) gastropod mollusk, (M) polychaete worm, (N) ammonoid cephalopod, and (O) belemnoid cephalopod.

From E Winson in W S McKerrow (ed), *The Ecology of Fossils*, Gerald Duckworth & Company Ltd

North America, *Costirhynchia* of the Middle Jurassic in Europe, and the Upper Jurassic (Portlandian and Volgian) form *Rhynchonella*. Examples of the terebratulids are *Epithyris* of the Middle Jurassic in Europe and *Somalithyris* of the Oxfordian in Somaliland.

Bryozoans are represented by the encrusting cyclostomes of the warm Jurassic seas and include such genera as *Entalophora, Spiropora* (a spiral-shaped colonial form), and *Idmonea*. Members of another order of bryozoa, the cheilostomes, first appeared in the Jurassic and rapidly became abundant.

The bivalves, or pelecypods, normally a slow-changing group of mollusks, showed rapid expansion during the Jurassic, adding a dozen new families. Jurassic assemblages tended to include the schizodonts, characterized by few and distinct hinge teeth, and dysodonts, distinguished by weak or absent teeth; included in the latter group are the pectinid and oysterlike forms. Heterodonts were relatively unimportant during this time. Oysters and pectinids, such as the genus *Lima*, with eulamellibranch gills (specialized gills in which the lamellae consist of solid sheets of tissue), continued to develop toward modern varieties. Pelecypods and gastropods, particularly the prosobranchs, were prominent in the shallow seas of the Jurassic geosynclines. The high-spired nerineids evolved during the Jurassic. Pulmonate gastropods, such as *Limonaea* and *Helisoma*, are found in lake beds of Purbeckian age, while others, like *Valvata*, are marine. In terms of varying shell morphology *Itieria*, a gastropod from the Upper Jurassic of France, has a complex umbilicate shell. Another group, the archaeogastropods, expanded markedly during the Late Jurassic as well.

Ammonoid index fossils

The ammonoid cephalopods, like the brachiopods, recovered remarkably from near extinction in Late Triassic time on the evolutionary success of two relict genera, *Phylloceras* and *Lytoceras*. Complex suture patterns—the trace of the chamber partitions or septa edges on the shell walls—and varying shell morphology and ornamentation provide the variety of index fossils for the more than 50 ammonoid zones that delineate Jurassic stratigraphy (Table 19). While the Lower Jurassic is characterized by tightly coiled ammonoid shells, loosely coiled forms appear in the Middle Jurassic. Upper Jurassic ammonoids are more distinctively ornamented, and Jurassic ammonoids in general are larger than their predecessors. Later varieties demonstrated regressive shell form, tending toward uncoiling and linear types. Only one group of nautiloid cephalopods survived the Late Triassic extinctions. Coleoid cephalopods—those lacking external skeletons like the belemnoids (or belemnites), are represented in the Jurassic fossil record by abundant mineralized phragmocones (conical internal shells).

Arthropods, mainly the tiny aquatic crustaceans of the subclass Ostracoda, figure prominently in the stratigraphic zonation of the Jurassic. Decapods, crabs, and the first lobsters, which appear in the Upper Jurassic of Europe, are found in Jurassic benthic communities. The modern king crab Xiphosura, commonly called the horseshoe crab, also originated in the Jurassic. Isopods and insects constitute the terrestrial forms. Some insect groups that are represented in the fossil record of the period include the Odonata (dragonflies), Coleoptera (beetles), Neuroptera (lacewings), Diptera (flies), and Hymenoptera (bees, ants, and wasps).

Prominent Jurassic echinoids include both stalked and unstalked varieties of pelmatozoans, such as the crinoids and the free-moving echinoids (sea urchins) and stelleroids. The Mesozoic crinoids are principally in the subclass Articulata, which arose in the Triassic. Stem-bearing forms of this group include *Isocrinus* and *Pentacrinites*. Holothuroid (sea cucumber) impressions and spicules, which take the form of wheels, crosses, and hooks, have been found in the Lias of Germany. *Ophioglypha* is a representative genus of Jurassic sea stars, while *Plesiocidaris* constitutes a representative fossil form of the so-called regular echinoids (those with radially symmetrical bodies). Irregular echinoids (those with bilaterally symmetrical bodies) first appeared in the Jurassic and include *Clypeus*, a European variety.

Zonation

Correlation between European and North American Jurassic strata has been facilitated through biostratigraphic zonations using primarily ammonoid index fossils. The ammonoid zones (Table 19), supplemented by other invertebrate and protozoan index fossils, have been used to identify the stages of the Jurassic rocks found throughout the western interior of North America, as well as in Alaska, Mexico, and Cuba.

Lower to Middle Jurassic ammonoid assemblages that indicate Toarcian and Bajocian ages are characterized by *Sonninia* and *Tmetoceras* and are found in the Kialagvik and Tuxedni formations in Alaska. *Defonticeras* and *Stemmatoceras* occur in the Twin Creek Limestone of Bajocian age in Wyoming, and *Arnioceras* dates the Barranca Formation of the Mexican Sonora as Lias (Early Jurassic) in age.

The macrocephalitid ammonite *Arcticoceras*, which marks the Callovian Stage in Greenland, is found in the Curtis Formation of Utah, the Rierdon Formation of Montana, and the Sundance and Carmel formations of Wyoming. *Arcticoceras* and *Gowericeras* occur in the Lower Callovian in the northern interior of North America, as well as in the Lower Sundance Formation of Wyoming. The *Cadoceras* fauna identifies the Shelikof Formation and the Chinitna Siltstone in Alaska as Callovian, which is the equivalent of the *Proplanuites* to *Erymoceras* zones in Europe. *Cardioceras* and its associated fauna signify the Oxfordian in both Europe and North America; *C. cardiforme* marks the lower Oxfordian, as, for example, in the Stump Sandstone and the Upper Sundance Formation of Utah and Wyoming. *Cardioceras*, together with *Goliathiceras* and *Pachycardioceras*, represent the Late Oxfordian in the Swift Formation of Montana and the Jagua Formation of Cuba. *Phylloceras* and other related ammonites form a zonal equivalent to the Oxfordian *Peltoceras* zone of central Europe. In the Gulf Coast of the United States, the Smackover Formation contains *Dichotomosphinctes* and an associated fauna considered to be Late Oxfordian in age; *Amoeboceras* in the Mariposa Slate in the Sierra Nevada range indicates an Oxfordian to Early Kimmeridgian age; and the presence of *Cardioceras*, along with *Amoeboceras*, in the Naknek Formation in the Cook Inlet area of Alaska suggests that it is of Oxfordian age.

In central and southern Mexico, ammonite zones show the presence of stages between the Bathonian and the Portlandian. Marls of Bathonian age are identified by a *Strenoceras* assemblage. The occurrence of *Idoceras* in the Zuloaga Limestone in southern Mexico and the Olvido Formation in the Sierra Madre Oriental, on the other hand, points to an Oxfordian origin. The Tamán Formation is shown to be Kimmeridgian through a *Haploceras* and *Aspidoceras* assemblage, and the Pimienta Formation is Portlandian, based on the *Parodontoceras* fauna. In the Alaskan Range, Oxfordian to Kimmeridgian beds contain *Aucella*, and Kimmeridgian to Portlandian rocks have *Aucella*, along with *Amoeboceras*. A similar fauna is found in the Kupreanof and Gravina islands in the Alaskan panhandle and in the Kingak Shale in the Canning River area of the Arctic slope.

Vertebrates. Chondrichthyes, mainly sharks, and Osteichthyes, the bony fish, including teleosts, ray-finned fish (or actinopterygians), and holosts (or ganoid fish), were the principal vertebrate swimmers of the Jurassic seas. The teleosts developed ossified vertebrae at this time and showed considerable change in bone structure, fins, and tail. The teleosts are the predecessors of the most prevalent modern fish. Early amphibian groups, such as the labyrinthodonts, became extinct by the Late Triassic and were succeeded by the first Anura (frogs and toads) and salamanders in the Jurassic.

The dominant land animals of the Jurassic were reptiles, the most significant of which belonged to the superorder known as Archosauria. The archosaurs included the thecodonts of the Triassic from which the dinosaurs descended, as well as the crocodiles and pterosaurs (flying reptiles).

The dinosaurs are divided into two principal groups on the basis of pelvic and hip structure: the saurischians and the ornithischians. The pelvis and hip of saurischians were reptilian (or lizardlike), whereas those of the ornithischians were birdlike. The sauropods, one of the basic types of saurischians, appeared in the Early Jurassic and became abundant in the Late Jurassic. They included both carnivorous and herbivorous forms. The latter were among the largest of the dinosaurs, with certain varieties reaching up to 30 metres in length. These gigantic vegetarians included the genera *Apatosaurus, Brachiosaurus,* and *Diplodocus.* (The *Apatosaurus* was long referred to as "Brontosaurus," which was in actuality a form created by workers who inadvertently placed the wrong skull on the body of an *Apatosaurus.*) Another basic saurischian type was the meat-eating theropod, an early representative of which was *Epanteria.* This genus was succeeded by *Tyrannosaurus* and the smaller *Allosaurus.* The apatosaurs presumably could fend off the allosaurs but not the other larger carnivores. The theropods were all bipedal; they had powerful hind legs and shortened forelegs with sharp claws.

The ornithopods, the first major suborder of ornithischians, appeared in the Late Jurassic and were much smaller, ostrichlike forms with large brains. They included the hadrosaurs and the anatosaurs, the bipedal duckbill forms formerly known as trachodons. The anatosaurs were hollow-boned and probably warm-blooded and may have been the ancestors of the birds. It is thought that other dinosaur groups may also have been warm-blooded.

Armoured dinosaurs emerged during the Jurassic. One notable variety that roamed North America during the Middle Jurassic was the stegosaur, a relatively large ornithischian type characterized by a double row of vertical bony plates along the dorsal midline and a spiked tail. (For a detailed treatment of the dinosaurs, see the article DINOSAURS.)

The reptiles of the Jurassic had diverse habitats. The plesiosaurs, with paddle-shaped limbs and generally long necks, shared the seas with the ichthyosaurs, highly specialized marine reptiles that have often been likened to such fast-swimming modern fish as tuna and marlin. Two other reptile groups, the crocodiles and marine turtles, are also known from the Early Jurassic, while the lizard made its appearance during Late Jurassic time. The pteropods, airborne flyers or perhaps gliders, were common throughout the Jurassic. The Jurassic pteropods were very small, comparable in size to sparrows and somewhat larger birds. They had hollow bones and were thought to be warm-blooded.

Early birds and mammals

Birds evolved in the Late Jurassic and closely resembled some small dinosaurs. Fossils of the first bird, the crow-sized *Archaeopteryx,* were found in the lithographic limestones at Solenhofen in Germany. This bird was very reptilian in appearance, having teeth, a vertebrae-supported tail, three claws at the wing tips, and scales. The specimens, however, have feathers. By Jurassic time, six orders of mammals—mainly small, shrewlike creatures—existed. Multituberculates, pantotheres, and triconodont mammals were important stock from which later forms developed. These early mammals have been classified on the basis of tooth morphology and were probably omnivorous in diet.

Plants. Jurassic plant communities differed considerably from their predecessors. The seed-fern floras of the early Mesozoic, like the *Glossopteris* flora of Gondwana, declined in importance. The cycads and cycadeoides, the palmlike gymnosperms, proliferated to the extent that the Jurassic has been called the "Age of Cycads." Another variety of gymnosperms, the conifers, made up a large component of Jurassic forests. These included araucarian pines, such as the Norfolk pine that exists today in the South Pacific. The ginkgo, a fruit-bearing gymnosperm, also was widespread during the Jurassic. It has been debated whether or not angiosperms, the flowering plants, evolved as early as the Jurassic; however, pollen of monocotyledonous angiosperms and palm stems have been reported in rocks of Jurassic age. (L.Si.)

CRETACEOUS PERIOD

General considerations. The Cretaceous, the third and final period of the Mesozoic Era, began about 144 million years ago and ended 66.4 million years ago (see Table 4). Spanning more than 77 million years, the Cretaceous is the longest of all the Phanerozoic periods; it represents more time than has elapsed since the last dinosaurs roamed the Earth. The rocks that were either deposited or formed during the Cretaceous Period make up the Cretaceous System.

The name Cretaceous is derived from *Creta,* the Latin word for chalk, and was first proposed by J.B.J. Omalius d'Halloy in 1822. D'Halloy had been commissioned to make a geologic map of France, and part of his task was to decide upon the geologic units to be represented by it. One of his rock units, the Terrain Crétacé, included chalks and underlying sands. Chalk is a soft, fine-grained type of limestone composed predominantly of the armour-like plates of planktonic coccolithophores, floating algae that flourished during the Late Cretaceous. Most Cretaceous rocks are not chalks, but most chalks were deposited during the Cretaceous.

Major subdivisions. The Cretaceous Period is divided into two nearly equal epochs, the Early Cretaceous and the Late Cretaceous. Each of these epochs is subdivided into six ages of quite variable duration, as shown in Table 20. The longest age, the Albian, lasted some 15.5 million years, more then 15 times the Coniacian, which lasted a mere 1 million years.

Stages

The definition of these Cretaceous ages was initiated during the mid- to late 1800s, when geologists working in France, Belgium, The Netherlands, and Switzerland recognized and named the 12 Cretaceous stages. The unit of time, age, is measured from the beginning of a given stage to the beginning of the succeeding stage. Each of the stages was defined by the rocks, sediments, and fossils at a particular locality. For example, A.D. d'Orbigny defined and described the Cenomanian Stage in 1847, based on some 847 species characteristic of the strata, and confirmed Le Mans, Fr., as the type area. The Cenomanian Age is defined on the basis of the rocks, sediments, and fossils in the type area for the Cenomanian Stage.

A type area is not always the best place to define a stage. The type area for the Coniacian Stage, for example, is in the environs of Cognac, Fr., but there the boundary with the underlying Turonian is marked by a discontinuity, and one stratigraphically important fossil group, the inoceramid bivalves, is poorly represented. These conditions make correlation of the base of the Coniacian Stage difficult at sites away from the type area.

Since the inception of the 12 Cretaceous stages, geologists have worked to solve such problems caused by incompleteness of the stratigraphic record and fossils of poor biostratigraphic utility in type areas. It is now customary to define the base of one stage and to consider that stage as continuing until the beginning of the next younger stage. Researchers meet periodically to discuss problems of stage boundaries and to suggest solutions. In 1983 a group of geologists from around the world met in Copenhagen and suggested that alternative type areas be designated for all the 11 stage boundaries discussed. Further, they suggested that the overly long Albian Stage be divided into three substages: the Lower, Middle, and Upper Albian. It was agreed at that symposium that stages are "packages of zones" and that the most sensible way to define a stage is by the base of the earliest biozone at a boundary type area. Traditionally, ammonites have been used to define biozones within the type area of Cretaceous stages, but other animals, such as inoceramid bivalves, belemnites, and even calpionellids, are sometimes used.

The number of usable biozones for the Cretaceous varies from area to area. For example, about 25 ammonite zones are employed in the type areas of western Europe for the whole of the Cretaceous, but at least 55 are recognized in the Upper Cretaceous alone for the western interior of North America.

The mid-Cretaceous is an informal subdivision of the Cretaceous and is usually taken to mean the Aptian, Albian, Cenomanian, Turonian, and Coniacian ages. A project called "Mid-Cretaceous Events" was undertaken from 1974 to 1985 and focused the efforts of several hundred geologists on that particular period of time. These investigators were particularly interested in the his-

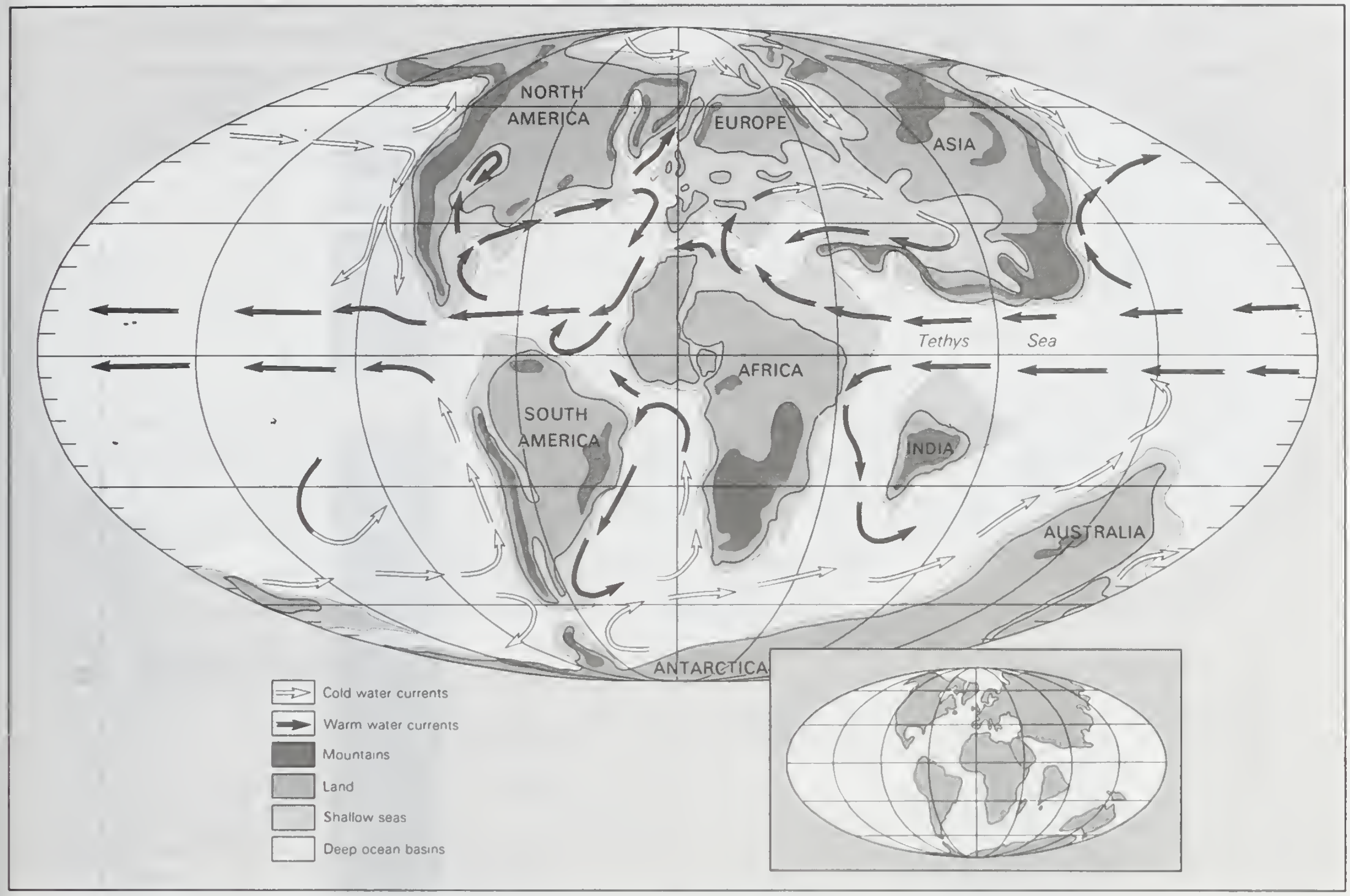


Figure 37: Distribution of landmasses, mountainous regions, shallow seas, and deep ocean basins during Late Cretaceous time. Included in the paleogeographic reconstruction are cold and warm ocean currents. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset at the lower right.

Adapted from C.R. Scotese, The University of Texas at Arlington

tory of major transgressions and regressions, the geologic evolution of the South Atlantic Ocean, "oxygen-deficient sedimentational processes" (mostly those associated with black shales), and biological issues, such as evolutionary sequences, biogeography, and paleoecology. Much of the information that follows is derived from the findings of this project and its participants.

Time of transition

Distinctive features. The Cretaceous Period has many distinctive features, partly because of its long duration and partly because it is a major part of the period of transition between the very different earlier Earth and the Cenozoic, a world relatively similar to that of the present day in terms of life-forms and the positions of continents. In addition, the Cretaceous is young enough so that many of the rocks which contain details about the period have not been deformed or eroded and are relatively close to the surface.

The Cretaceous opened with the landmasses assembled essentially into two continents, Laurasia in the north and Gondwana in the south, nearly separated by the equatorial Tethys seaway. (As was noted elsewhere, the various segments of Laurasia and Gondwana had already started to rift apart by this time. North America had just begun pulling away from Eurasia during the Jurassic, and South America had started to split off from Africa, from which India, Australia, and Antarctica were also in the process of separating.) The period closed with most of the present-day continents separated from each other by such expanses of water as the North Atlantic and South Atlantic. At the end of the Cretaceous, India was adrift in the Indian Ocean, and Australia was still connected to Antarctica. Figure 37 shows the positions of the landmasses for the Maastrichtian, about 4 million years before the close of the period.

Sea level was high throughout much of the Cretaceous Period. At its maximum height, only about 18 percent of the Earth's surface remained uncovered by water, as compared to approximately 28 percent today. At times, Arctic waters were connected to the Tethys seaway through the middle of North America and the central portion of Russia. Marine animals living in the South Atlantic had a seaway for migration to the Tethys via what is presently Nigeria, Niger, Chad, and Libya on several occasions during the Cretaceous. Most of western Europe, eastern Australia, parts of Africa, South America, India, Madagascar, Borneo, and certain other areas that are now land were entirely covered by marine waters for some interval of Cretaceous time.

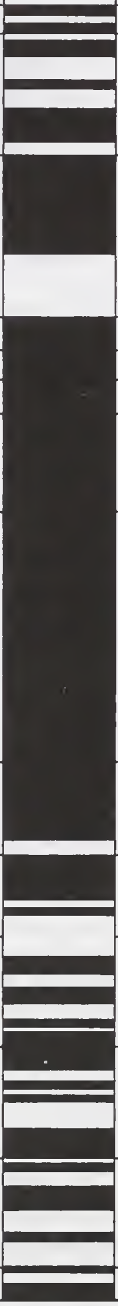
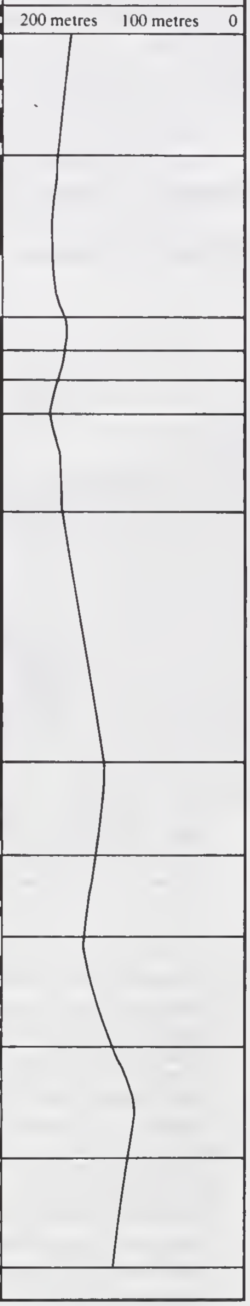
Water circulation and mixing was not as great as it is today because most of the oceans (*e.g.*, the developing North Atlantic) were constricted, and the temperature differences between the poles and the equator were minimal. Thus the oceans experienced frequent periods of anoxic bottom water that reveal themselves today as black shales. Sometimes, particularly during the mid-Cretaceous, oxygen-deficient conditions extended to the waters covering the present-day land areas, the so-called epicontinental seas, as attested to by the black shales in the western interior of North America.

The Late Cretaceous was a time of great productivity in the world's oceans, as borne out by the deposition of thick beds of chalk in western Europe, eastern Russia, southern Scandinavia, the Gulf Coast of North America, and Western Australia.

The Cretaceous was magnetically quiet relative to the subsequent Tertiary Period. In fact, magnetic reversals are not noted from the early Aptian to the late Santonian, a period of some 34 million years. Table 20 shows the Cre-

Table 20: Stratigraphic Subdivisions of the Cretaceous Period*

* Temporal relationships of sea level, paleomagnetic reversals, and rocks of representative stable areas are shown.

period	epoch	age	age (millions of years ago)	"classic" type area	proposed base defining biozones	magnetic polarity	worldwide sea-level curve
Cenozoic			66.4				
Cretaceous	Late	Maastrichtian	74.5	Maastricht, S.E. Netherlands	*Belemnella lanceolata*		
		Campanian	84.0	Grande et Petite, Champagne, France	*Gonioteuthis granulataquadrata*		
		Santonian	87.5	Saintes, W. France	*Texanites* (*Texanites*)		
		Coniacian	88.5	Cognac, W. France	*Forresteria petrocoriensis*		
		Turonian	91.0	Tours, W. France	*Pseudaspidoceras flexosum*		
		Cenomanian	97.5	Le Mans, France (Paris Basin)	*Hypoturrilites schneegansi*		
	Early	Albian	113	Aube, France (Paris Basin)	*Dipoloceras cristatum* *Lyelliceras lyelli* *Leymeriella schammeni*		
		Aptian	119	Apt, S.E. France	*Prodeshayites*		
		Barremian	124	Angles, France	*Pseudothurmannia*		
		Hauterivian	131	Hauterive, Switzerland	*Acanthodiscus*		
		Valanginian	138	near Valangin, Switzerland	*Thurmanniceras otopeta*		
		Berriasian	144	Berrias, S. France	not discussed		
Jurassic							

taceous magnetic reversals. The Earth's synodic months have changed regularly for at least the last 600 million years owing to tidal friction and other forces that slow up the Earth's rotation. The rate of change in the synodic month was minimal for most of the Cretaceous but has accelerated since. The reasons for these two anomalies are not well understood.

Major oil deposits

Economic significance. In approximately 30 million years during the middle of the Cretaceous, more than 50 percent of the world's known petroleum reserves were formed. Almost three-fourths of this mid-Cretaceous petroleum accumulated in a relatively small region around the Persian Gulf. Much of the remainder accumulated in another limited region of the Americas between the Gulf of Mexico and Venezuela. Evidently the low-latitude Tethys seaway collected large amounts of organic matter along its margins, which today is found as petroleum in the Gulf Coast of the United States and Mexico, the Maracaibo Basin in Venezuela, the Surt (or Sirte) Basin in Libya, and the Persian Gulf region.

Other mineral deposits of commercial value occur in the circum-Pacific mountain systems and chain of island arcs. Such metals as gold, silver, copper, lead, zinc, molybdenum, tungsten, tin, iron, and manganese were concentrated into ore deposits of various dimensions during episodes of igneous activity in the late Mesozoic.

Cretaceous rocks. *Occurrence and distribution.* The occurrence and distribution of Cretaceous rocks resulted from the interplay of many forces. The most important of these are the position of the continental landmasses, level of the sea relative to these landmasses, local tectonic and orogenic activity, climatic conditions, availability of source material (for example, sands, clays, and even the remains of marine animals and plants), igneous activity, and the history of the rocks and sediments after intrusion or deposition. Many Cretaceous sedimentary rocks

Table 20: (continued)

North American western interior basin centre	North American Gulf Coast—Texas	eastern England	western Libya	southeast Nigeria	Australia—Great Artesian Basin
Fox Hill Sandstone	Navarro Group marl and clay	Trimingham Chalk	Lower Tar Marl member— Zmam Formation	Nsukka Formation sandstone and shale	no rocks
Pierre Shale	Taylor Marl and Anacacho Limestone	Norwich Chalk	Mizda Formation marls and limestones	Ajali Sandstone	Winton nonmarine sandstone, shale, and coal
Eagle Sandstone	Austin Chalk	Upper Chalk	Gharian Limestone Formation	Mamu Formation sandstone and shale	Tambo marl and clay (marine)
Niobrara Formation, chalk	Eagleford Shale	Middle Chalk	Jefren Marl Formation	Nkpuro Shale	Styx coal
Carlile Shale Greenhorn Limestone	Buda Limestone Del Rio Clay Georgetown Limestone	Lower Chalk	Ain Tobi Limestone Formation	Awgu Shale	Roma marl and clay (marine)
Graneros Shale	Fredericksburg and Trinity groups limestone and clay	Upper Greensand	Uazzen Formation dolomite	Eze-Aku Sandstone	Bathesdale Formation mainly nonmarine, partly brackish water sediments
Dakota Sandstone sand, shale, and conglomerate		Gault Clay	Giado Formation sandstone and shale	Odukpani Formation sandstone, shale, and limestone	
		Lower Greensand	Chicla Sandstone Formation	Asu River Shales	
		Wealden sandstones and clays	Cabao Sandstone Formation		
		Darlston Beds			

have been eroded since their deposition, while others are merely covered by younger sediments or are presently underwater or both.

Figure 37 shows the position of the present-day continental landmasses during the Maastrichtian approximately 70 million years ago. At the very beginning of the Mesozoic these landmasses had been together. As was mentioned earlier, South America, Africa (including the adjoining pieces of what are now the Arabian Peninsula and Middle East), Antarctica, Australia, India, Madagascar, and several smaller landmasses were joined as Gondwana in the south, while North America, Greenland, and Eurasia (including Southeast Asia) formed Laurasia in the north. The breakup of the supercontinent of Pangaea, which had started more than 100 million years earlier during the Early Jurassic, showed major developments in the Cretaceous. Africa split from South America, the last land connection being that between Brazil and Nigeria. This separation was complete by about Aptian–Albian time, resulting in the joining of the South Atlantic Ocean with the widening North Atlantic. In the region of the Indian Ocean, Africa and Madagascar separated from India, Australia, and Antarctica in Late Jurassic to Early Cretaceous times. Once separated from Australia and Antarctica, India began its journey northward, which culminated in a collision with Eurasia during the Cenozoic (see below the section *Cenozoic Era*). Madagascar broke away from Africa during the later Cretaceous. During the Late Cretaceous, Greenland separated from North America. A graphic example of the influence of continental fragmentation on the Cretaceous rock record can be seen in the stratigraphic column for southeastern Nigeria in Table 20. The stratigraphic record begins in the Albian only after the South Atlantic opened.

High sea levels

Sea level was higher during most of the Cretaceous than at any other time in Earth history. In general, the world oceans were about 100 to 200 metres higher in the Early Cretaceous and roughly 200 to 250 metres higher in the Late Cretaceous than at present. The high Cretaceous sea level is thought to have been primarily the result of water in the ocean basins being displaced by the enlargement of the mid-oceanic ridges. Table 20 gives the overall trend in sea level, but minor peaks and troughs are known to have occurred. The rocks shown in the Table illustrate well the effect of the high Cretaceous sea level. Chalks and limestones, for example, were deposited in the western interior of North America only during the early Late Cretaceous, when sea levels were at their highest.

As a result of higher sea levels during the Late Cretaceous, marine waters inundated the continents, creating relatively shallow epicontinental seas in North America, South America, Europe, Russia, Africa, and Australia. In addition, all continents experienced diminution of land area adjacent to the major oceans.

The effects of these higher sea levels were not felt to the same extent by each continent, because the various continents experience movement of their crustal level due to isostasy and tectonism. For example, if a continent is emergent due to isostatic rebound, the vertical movement could exceed the rise in sea level and so the continent would not experience transgression of marine water but rather regression. Table 21 shows the records of transgression and regression during the Cretaceous Period in selected platforms of the world. When a platform is underwater, sedimentation occurs; when it is not, erosion takes place. The rock record for the Great Artesian Basin in east-central Australia (see Table 20), for instance, shows marine rocks for much of the Aptian–Albian but nonmarine sediments during the Cretaceous maximum transgression near the end of the Cenomanian.

A comparison of the rock record for the North American western interior with that for eastern England (see Table 20) reveals chalk deposition in eastern England from Cenomanian to Maastrichtian time, but chalks and marine limestone are limited to late Cenomanian through early Santonian time in North America. The two areas have nearly identical histories of transgression, as indicated in Table 21. It has been noted that the land areas of western Europe during the Late Cretaceous were limited to a few stable regions that represent low-lying islands within a chalk sea. Furthermore, sedimentological evidence indicates an arid climate that would minimize erosion of these islands and limit the input of sands and clays into the basin. In contrast, the North American western interior was receiving abundant clastic sediments that were being eroded from the new mountains along its western margin created by the Sevier orogeny of Cretaceous time.

In addition to the areas that have been mentioned above, Cretaceous rocks crop out in the Arctic, Greenland, central California, the Gulf and Atlantic coastal plains of the United States, central and southern Mexico, and the Caribbean islands of Jamaica, Puerto Rico, Cuba, and Hispaniola. In Central and South America, Cretaceous rocks are found in Panama, Venezuela, Colombia, Ecuador, Peru, eastern and northeastern Brazil, and central and southern Argentina. Most European countries have Cretaceous rocks exposed at the surface. North Africa, West Africa, coastal South Africa, Madagascar, Arabia, Iran, and the Caucasus all have extensive Cretaceous outcrops, as do eastern Siberia, Tibet, India, China, Japan, Southeast Asia, New Guinea, Borneo, Australia, New Zealand, and Antarctica.

Types. The rocks and sediments of the Cretaceous System show considerable variation in their lithologic character and the thickness of their sequences. Mountain-building episodes accompanied by volcanism and plutonic intrusion took place in the circum-Pacific region and in the area of the present-day Alps. The erosion of these mountains produced clastic sediments, such as conglomerates, sandstones, and shales, on their flanks. The igneous rocks of Cretaceous age in the circum-Pacific area are widely exposed.

The Cretaceous Period was a time of great inundation by shallow seas that created swamp conditions favourable for the accumulation of fossil fuels at the margin of land areas. Coal-bearing strata are found in some parts of Cretaceous sequences in Siberia, Australia, New Zealand, Mexico, and the western United States.

Farther offshore, chalks are widely distributed in the Late Cretaceous. Another rock type called the "Urgonian" limestone is similarly widespread in the Upper Barremian–Lower Aptian. This massive limestone facies, whose name is commonly associated with rudists (a reef-building bivalve of the Mesozoic), is found in Mexico, Spain, southern France, Switzerland, Bulgaria, the southern Soviet Union, and North Africa.

Formation of black shales

The mid-Cretaceous was a time of extensive deposition of carbon-rich shale with few or no benthic fossils. These so-called black shales result when there is severe deficiency of oxygen in the bottom waters of the oceans. Poor ocean circulation is suggested as the cause, and the poor circulation is thought to have resulted from the generally warmer climate that prevailed during the Cretaceous, the much smaller than present temperature difference between the poles and the equator, and the restriction of the North Atlantic, South Atlantic, and Tethys. Cretaceous black shales are extensively distributed on various continental areas, such as the western interior of North America, the Alps, the Apennines of Italy, western South America, Western Australia, western Africa, and southern Greenland. They also occur in the Atlantic Ocean, as revealed by the Deep Sea Drilling Program (a scientific program initiated in 1968 to study the ocean bottom), and in the Pacific, as noted on several seamounts.

In North America the Nevadan orogeny took place in the Sierra Nevada and Klamath Mountains from Late Jurassic to Early Cretaceous times; the Sevier orogeny produced mountains in Utah and Idaho in the mid-Cretaceous; and the Laramide orogeny, with its thrust faulting, gave rise to the Rocky Mountains and Sierra Madre Oriental during the Late Cretaceous to Early Tertiary. In the South American Andean system, mountain building reached its climax in mid-Late Cretaceous. In Japan the Sakawa orogeny proceeded through a number of phases during the Cretaceous.

In typical examples of circum-Pacific orogenic systems, regional metamorphism of the high-temperature type and large-scale granitic emplacement occurred on the inner continental side, whereas sinking, rapid sedimentation, and regional metamorphism predominated on the outer oceanic side. The intrusion of granitic rocks, accompanied in some areas by extrusion of volcanic rocks, had a profound effect on geologic history. This is exemplified by the upheaval of the Sierra Nevada, with the intermittent emplacement of granitic bodies and the deposition of thick units of Cretaceous shales and sandstones with many conglomerate tongues in the Great Valley of California.

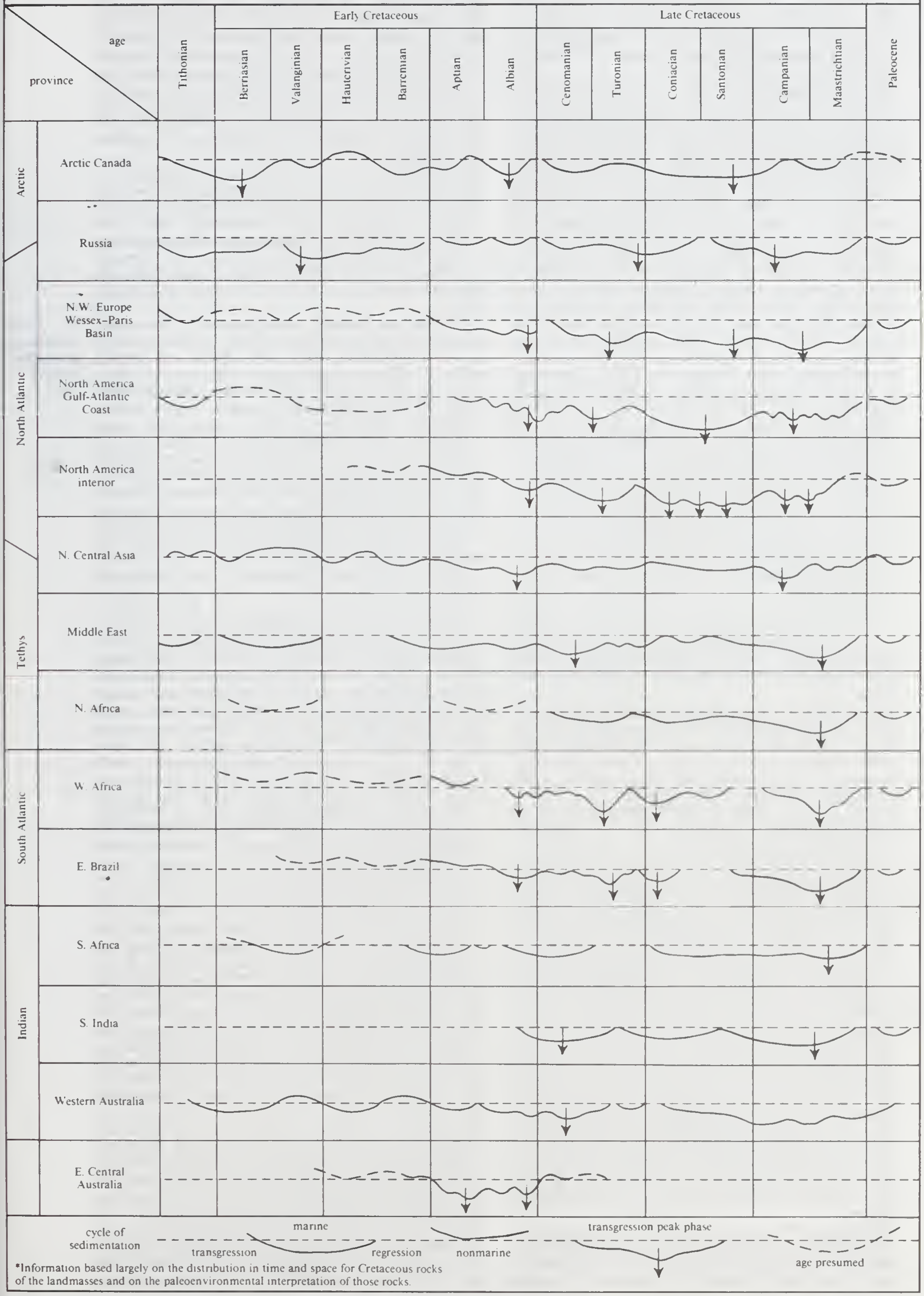
Volcanic seamounts of basaltic rock with summit depths of 1,300 to 2,100 metres are found in the central and western Pacific. Some of them are flat-topped, with shelves on their flanks on which reef deposits or gravels accumulated, indicating a shallow-water environment. Some of the deposits contain recognizable Cretaceous fossils. Although the seamounts were formed at various times during the late Mesozoic and Cenozoic eras, a large number of them were submarine volcanoes that built up to the sea surface during the Cretaceous. They sank to their present deep levels some time after the age indicated by their youngest shallow-water fossil.

In west-central India, the Deccan traps consist of more than 1,200 metres of basaltic lava flows that erupted from the Late Cretaceous to the Eocene Epoch of the Tertiary over an area of some 500,000 square kilometres. Volcanic activity on the western margin of the North American epicontinental sea frequently produced ashfalls over much of the western interior seaways. One of these, the "X" bentonite near the end of the Cenomanian, can be traced more than 2,000 kilometres from central Manitoba to north Texas.

Ammonite index fossils

Correlation. Correlation of Cretaceous rocks is usually accomplished using fossils. Ammonites are the most widely employed fossils in terms of both frequency of use and geographic extent, but no single fossil group is capable of worldwide correlation of all sedimentary rocks. Most

Table 21: Cretaceous Record for Transgressions and Regressions of Marine Waters onto Selected Landmasses*



*Information based largely on the distribution in time and space for Cretaceous rocks of the landmasses and on the paleoenvironmental interpretation of those rocks.

ammonites, for example, did not occur in all latitudes because some preferred the warmer waters of the Tethys seaway, while others resided in cooler boreal waters. Furthermore, ammonites are rarely found in sediments deposited in nonmarine and brackish environments, and they are seldom retrieved from boreholes sufficiently intact for confident identification.

Many ammonites are very good index fossils, but they are not perfect. When Cretaceous stage boundaries were proposed by the above-mentioned international group of geologists in Copenhagen in 1983, the problems of correlating the boundary between the Campanian Stage and the underlying Santonian were examined. The ammonite *Placenticeras bidorsatum,* the index species to the oldest Campanian zone of the "classic" zonation, is extremely rare in the type area of western France and appears to be restricted to northwestern Europe. Other ammonites discussed as possible alternatives were *Submortaniceras,* a genus known from Spain, the Natal province of South Africa, Madagascar, Mexico, and the Gulf and Pacific coasts of the United States and British Columbia but not from the type area, and the *Scaphites hippocrepis,* a species described from North America. In Europe, *Scaphites aquisgranensis,* a late form of *S. hippocrepis,* occurs together with *P. bidorsatum.* This *S. hippocrepis* form occurs widely in the Gulf and Atlantic coastal plains of North America, western Germany, Belgium, and The Netherlands, as well as in the North American western interior and the Aquitane basin of France. The belemnite species *Gonioteuthis granulataquadrata* is widely used in western Germany for definition of the boundary, but it has a restricted boreal distribution (see below *Cretaceous environment*). The extinction level of the free-living crinoid *Marsupites testudinarius,* the appearance of the coccolith *Aspidolithus parcus,* and the first occurrence of the foraminiferans *Bolivinoides strigillatus* and *Globotruncana arca* are all used to define the boundary in some regions but not in others.

It was generally agreed that a boundary level close to the currently used appearance of the belemnite species *G. granulataquadrata* in the boreal realm—*i.e.,* temperate paleobiogeographic region—would be desirable because this boundary could be correlated with a number of other events.

It is desirable to have a reference section for the boundaries of all Cretaceous stages, and the Campanian example above serves to illustrate the variety of fossil groups used to define boundaries and the complexity of the definition problem. The boundaries of the other stages have similar problems of restricted distribution for fossils in the classic type areas. Other fossil types useful for defining Cretaceous stage boundaries are inoceramid bivalves, echinoids, larger foraminiferans, and calpionellids.

On a more local scale, correlation can be achieved using a variety of fossil groups. Rudist, inoceramid, and exogyrid bivalves have been used in many areas to subdivide or zone the Cretaceous Period for the purpose of correlation. Rudist bivalves, for example, have been employed in conjunction with larger foraminiferans to zone sediments of the Tethyan regions of France and Yugoslavia. Echinoids and belemnites have been used together to zone the Late Cretaceous of eastern England. Angiosperm pollen provides for recognition of zones for the Late Cretaceous of the North American Atlantic Coastal Plain.

Use of angiosperm pollen

Some fossil groups are useful for correlation between several regions because of their nektonic or planktonic life habit. Principal among these are ammonites, belemnites, planktonic foraminiferans, calcareous nannofossils, and radiolarians. In North America, for instance, Late Cretaceous strata in Texas, Arkansas, Mexico, and the Caribbean have been correlated using planktonic foraminiferans. Occasionally ostracods (small bivalved crustaceans) are useful; *e.g.,* they have been used to correlate Early Cretaceous strata of northwestern Europe with those of the Russian Platform.

The epicontinental sea of the North American western interior has been particularly well studied, primarily because it can be zoned to great precision. Sixty ammonite zones, to cite a case in point, are recognized in the rocks deposited between the late Albian and the late Maastrichtian. In addition, frequent bentonite beds resulting from the volcanic ash of the Sevier orogenic events provide radiometric dates with which to independently verify the synchronicity of the ammonite zones. This detailed resolution of about 0.5 million years per zone is unusual for the Cretaceous Period. Interestingly, the youngest Cretaceous biozone of the North American western interior is recognized regionally by the occurrence of the dinosaur genus *Triceratops,* because the last approximately one million years in that area are characterized by nonmarine sediments.

For some of the geologic record, more detailed subdivisions within zones can be developed on the basis of magnetic reversals. The Cretaceous Period, however, has a dearth of magnetic reversals. Specifically, only 16 reversals are noted for latest Jurassic to Aptian time, none for Aptian to late Santonian time, and just 9 from the late Santonian to the Cenozoic boundary (see Table 20). Magnetic reversals occur far more frequently in Cenozoic rocks.

Cretaceous environment. *Paleogeography.* The major geographic subdivisions of the world for the Cretaceous Period are the northern boreal, southern boreal, and Tethyan regions. The Tethyan region separates the other two and is recognized by the presence of reef-forming rudist bivalves, corals, larger foraminiferans, and certain ammonites that inhabited only the warmer Tethyan waters. Early in the Cretaceous, North America and South America separated sufficiently for the marine connection between the Tethys and Pacific to deepen substantially. The Tethys to Pacific marine connection allowed for a strong westward-flowing current, which is inferred from faunal patterns. For example, as the Cretaceous progressed, the similarity between rudist bivalves of the Caribbean and western Europe decreased, while some Caribbean forms have been found on Pacific seamounts, in Southeast Asia, and possibly in the Balkans.

Cretaceous geographic subdivisions

The Cretaceous of the northern boreal realm in North America, Europe, Russia, and Japan has been extensively studied because of many years of geologic research and the generally wealthier political entities that exist in those areas. It is known, for instance, that sediments in the southwestern Netherlands indicate several temperature swings for the Late Cretaceous. These temperature swings imply that the paleogeographic boundary between the northern boreal areas and the Tethys was not constant with time. Russian workers recognized six paleobiogeographic zones: boreal, which in this situation is equivalent to Arctic; European region; Mediterranean region, including the central Asian province; Pacific region; and two paleofloristic zonations of land. At present, studies of the southern boreal areas and the rocks representing the southern Tethys margin lack this level of detail.

As pointed out earlier, the positions of the landmasses changed significantly during the Cretaceous. This is not unexpected considering that the period lasted more than 77 million years. At the onset there existed the two supercontinents of Gondwana and Laurasia, which were barely attached at the junction of North and South America. When these enormous landmasses divided, the South Atlantic Ocean, Indian Ocean, Gulf of Mexico, and Caribbean came into being. By the end of the Cretaceous, the present-day continents were separate entities except for Australia, which was still joined to Antarctica. Also, India had not yet fused to Asia. The positions of the various continents were very nearly those shown in Figure 37 for the Maastrichtian shortly before the close of the Cretaceous.

Very high sea stand was a major factor influencing the paleogeography of the Cretaceous. In general, the maximum was lower in the Early Cretaceous than in the Late Cretaceous, but detailed study indicates from 5 to 15 different episodes of rises and falls in sea level. The high sea stands were not expressed equally in all regions.

Some regions were especially tectonically active during the Cretaceous. The Pacific margin of Canada, for one, shows evidence of an Early Cretaceous transgression, but by the Late Cretaceous much of the region had been uplifted some 800 to 2,000 metres. Japan also was very tectonically active, giving it a Cretaceous sedimentary record that varies from north to south, island to island, and time to time.

The history of sea-level change for the stable platforms is summarized in Table 21. Although the patterns are quite similar for many of the platform areas, several differences are notable. During the Berriasian, Valanginian, Hauterivian, and Barremian, parts of Arctic Canada, Russia, and Western Australia were underwater, but most of the other platforms were not. During the Aptian and Albian, east-central Australia experienced major transgressions, but those of the Late Cretaceous are not recorded there.

In the Late Cretaceous, most continental landmasses were transgressed but not always at the same time. One suggestion for the lack of a synchronous record is the concept of geoidal eustacy. It has been suggested based on observed patterns of Cretaceous rocks and physical calculations that, as the Earth's continents move about, the oceans bulge out at some places to compensate, and thus sea level rise is different from ocean basin to ocean basin.

Warm climatic conditions

Paleoclimate. In general, the climate of the Cretaceous Period was much warmer than at present, perhaps the warmest on a worldwide basis than at any other time during the Phanerozoic. The climate was also more equable in that the temperature difference from the poles to the equator was about one-half the present gradient. Floral evidence suggests that tropical to subtropical conditions existed as far north as 45°, and temperate conditions extended to the poles.

Temperatures were lower at the beginning of the period, rising to a maximum in the late Albian and then declining slightly with time until the decline was accentuated during the Campanian and Maastrichtian. Ice sheets and glaciers were almost entirely absent except in the high mountains. The late Maastrichtian was the coolest part of the Cretaceous, but the temperature was still much warmer than it is today.

Models of the Earth's climate for the mid-Cretaceous based on the positions of the continents, location of water bodies, and topography suggest that winds were weaker than at present. Westerly winds were dominant in the lower to mid-latitudes of the Pacific for the entire year. Winds were, however, westerly in the North Atlantic during winter but easterly during summer. Surface water temperatures were about 30° C at the equator year round, but 14° C in winter and 17° C in summer at the poles. A temperature of 17° C is suggested for the ocean bottom during the Albian but may have declined to 10° C by the Maastrichtian. These temperature values have been calculated from oxygen-isotope measurements of the calcitic remains of belemnites, planktonic foraminiferans, and benthic foraminiferans. These temperatures support models that suggest diminished ocean circulation both vertically and latitudinally. Such circulation patterns could account for the periods of black shale deposition during the Cretaceous.

In general, the Cretaceous would rank as the most arid period of the Phanerozoic. Not all scientists agree with this assessment, however. Evaporites are plentiful in the Early Cretaceous, a fact that seems to point to an arid climate. Yet, this situation may result more from constricted ocean basins than from climatic effects. The occurrence of evaporites mainly between 10° and 30° latitude suggests arid subtropics, but the presence of coals poleward of 30° indicates humid mid-latitudes. Occurrences of early Cretaceous bauxite and laterite, which are products of deep weathering in warm climates with seasonal rainfall, support the notion of humid mid-latitudes.

Other paleontological indicators suggest details of ocean circulation. The occurrence of early and mid-Cretaceous rudists and larger Tethyan foraminiferans in Japan may very well mean that there was a warm and northward-flowing current in the region. A similar occurrence of these organisms in Aptian–Albian sediments as far south as southern Tanzania seems to indicate a southward-flowing current along the east coast of Africa. The fact that certain warmwater life-forms that existed in the area of present-day Argentina are absent from the west coast of Africa suggests a counterclockwise gyre in the South Atlantic. In addition, the presence of larger foraminiferans in Newfoundland and Ireland points to the development of a "proto-Gulf Stream" by the mid-Cretaceous.

Cretaceous life. The lengthy Cretaceous Period constitutes a major portion of the interval of transition between ancient life-forms and those forms that dominate the Earth today. Many of these modern animals and plants, as, for example, the placental mammals and angiosperms, made their first appearance during the Cretaceous. (Some authorities maintain that certain varieties of angiosperms belonging to the class Monocotyledoneae evolved in the Late Jurassic based on fossil evidence.) Other groups, such as clams and snails, snakes and lizards, and teleost fishes (the ray-finned variety considered to be the most advanced of the bony fishes) developed distinctively modern characteristics by the end of the Maastrichtian.

Marine life

Characteristic fauna and flora. The marine realm can be divided into two paleobiogeographic regions, the Tethys and the boreal. The division is based on the occurrence of rudist-dominated organic reeflike structures. Rudists were large, rather unusual bivalves that had one valve shaped like a cylindrical vase and another that resembled a flattened cap. The rudists were generally dominant over the corals as framework builders. They rarely existed outside the Tethyan region, and the few varieties found elsewhere did not create reeflike structures. Rudist reeflike structures of Cretaceous age serve as reservoir rocks for petroleum in Mexico, Venezuela, and the Middle East.

Other organisms that were almost entirely restricted to the Tethys region were actaeonellid and nerineid snails, colonial corals, calcareous algae, larger benthic foraminiferans, and certain kinds of ammonites and echinoids. In contrast, belemnites were apparently confined to the colder boreal waters. Important bivalve constituents of the Cretaceous boreal marine biota were the reclining forms (*e.g., Exogyra* and *Gryphaea*) and the inoceramids, which were particularly widespread and therefore useful for biostratigraphic zonation.

Ammonites were numerous and were represented by a variety of forms ranging from the more usual coiled types to straight forms. Some of the more unusual ammonites, called heteromorphs, were shaped like fat corkscrews and fat hairpins. Such aberrant forms most certainly had difficulty moving about. Ammonites preyed on other nektonic and benthic invertebrates and were themselves prey to many larger animals, including the marine reptiles called mosasaurs.

Other marine reptiles were the long-necked plesiosaurs and more fishlike ichthyosaurs. Sharks and rays also were marine predators, as were the teleosts. One Cretaceous fish, *Xiphactinus,* grew to more than 4.5 metres and is the largest known teleost.

Flying reptiles and birds

In the air, the flying reptiles called pterosaurs dominated. One pterosaur from the latest Cretaceous of what is now Texas, *Quetzalcoatlus,* had a wingspan of about 15 metres. While it has been determined that birds developed from a reptilian ancestor during the Jurassic and Cretaceous, the fossil record for birds is too sparse to accurately document their evolution. *Hesperornis* was a genus of Cretaceous flightless, diving bird that had large feet and sharp backward-directed teeth adapted for preying on fish.

Although the fossil record is irregular in quality and quantity during the Early Cretaceous, it is obvious that dinosaurs continued their lengthy dominance on land. The Late Cretaceous record is much more complete, particularly in the case of North America and Asia. It is known, for instance, that during the Late Cretaceous many dinosaur types lived in relationships not unlike the present-day terrestrial mammal communities. Although the larger dinosaurs such as the carnivorous *Tyrannosaurus* and the herbivorous *Iguanodon* are the best known, many smaller forms also lived in Cretaceous times. *Triceratops,* a large three-horned dinosaur, inhabited western North America during the Maastrichtian age.

The land plants of the Early Cretaceous were similar to those of the Jurassic. They included the cycads, ginkgoes, conifers, and ferns. The angiosperms appeared by the Barremian, became common by the end of the Albian, and came to represent the major component of the terrestrial flora by the mid-Late Cretaceous. This flora included figs, magnolias, poplars, willows, sycamores, and herbaceous plants. With the advent of many new plant types, insects also diversified.

Emergence of placental mammals and marsupials

Forerunners of modern life-forms. Various types of small mammals that are now extinct existed during the Triassic and Jurassic. On the other hand, two important groups of modern mammals evolved during the Cretaceous. Placental mammals, which include most modern mammals (*e.g.*, rodents, cats, dogs, cows, pigs, and primates), evolved during the Late Cretaceous. Although almost always smaller than present-day rabbits, the Cretaceous placentals were poised to take over the terrestrial environments as soon as the dinosaurs vanished. Another mammal group, the marsupials, evolved during the Cretaceous as well. This group includes the native species of Australia, such as kangaroos and koalas, and the North American opossum.

Marine plankton took on a distinctly modern appearance by the end of the Cretaceous. The coccolithophores became so abundant in the Late Cretaceous that vast quantities accumulated to form the substance for which the Cretaceous Period was named—chalk. The planktonic foraminiferans also contributed greatly to fine-grained calcareous sediments. Less abundant but important single-celled animals and plants of the Cretaceous include the diatoms, radiolarians, and dinoflagellates. Other significant marine forms of minute size were the ostracods and calpionellids.

Mass extinctions. At or very close to the end of the Cretaceous, many animals that were important elements of the Mesozoic world became extinct. On land the dinosaurs perished, but the plant life was little affected. Of the planktonic marine flora and fauna, only about 13 percent of the coccolithophore and planktonic foraminiferan genera survived the extinction event (or events). Ammonites and belemnites became extinct, as did such marine nektonic reptiles as ichthyosaurs, mosasaurs, and plesiosaurs. Among the marine benthos, the larger foraminiferans (orbitoids) died out, and the hermatypic corals were reduced to about one-fifth of their genera. Rudist bivalves disappeared, as did bivalves with a reclining life habit such as the *Exogyra* and *Gryphaea*. The stratigraphically important inoceramids also died out. With the extinction of these animals, modern life-forms diversified and came to dominate the Earth.

Many theories have been proposed to explain the Late Cretaceous mass extinction. Since the early 1980s, much attention has been focused on the so-called asteroid theory formulated by the American scientists Walter and Luis Alvarez. In brief, this theory states that the impact of an asteroid (or meteorite) on the Earth may have triggered the extinction event by ejecting a huge quantity of rock debris into the atmosphere, enshrouding the Earth in darkness for several months or longer. With no sunlight able to penetrate this global dust cloud, photosynthesis ceased, resulting in the death of green plants and the disruption of the food chain. (For more detailed information on this hypothesis, see DINOSAURS: *Extinction.*)

The asteroid theory has met with considerable skepticism among paleontologists, many of whom prefer to look to environmental factors as the underlying mechanism for changes in biota. It has been noted that tectonic plate movements caused a major rearrangement of the world's landmasses, particularly during the latter part of the Cretaceous. The climatic changes resulting from such continental drift could have effected a gradual deterioration of habitats favourable to the dinosaurs and other animal groups that suffered extinction. It is, of course, possible that sudden catastrophic phenomena like an asteroid impact could have contributed to this environmental deterioration that brought about the demise of numerous faunal groups. (C.F.K.)

Cenozoic Era

The Cenozoic began about 66.4 million years ago and extends to the present (see Table 4). The term, originally spelled Kainozoic, was introduced by John Phillips in an 1840 *Penny Cyclopaedia* article to designate the most recent of the three major subdivisions of the Phanerozoic. Derived from the Greek for recent life, it reflects the sequential development and diversification of life on Earth from the Paleozoic (ancient life) through the Mesozoic (middle life). Today, the Cenozoic is internationally accepted as the youngest of the three subdivisions of the fossiliferous part of Earth history.

The Cenozoic Era is generally divided into two periods, the Tertiary and the Quaternary. The designations Tertiary and Quaternary, however, are relics of early attempts in the late 18th century at formulating a stratigraphic classification that included the now wholly obsolete terms Primary and Secondary. In 1856 Moritz Hörnes introduced the terms Paleogene and Neogene, the latter encompassing rocks equivalent to those described by Charles Lyell as Miocene and older and newer Pliocene (which included what he later called the Pleistocene; see above *Completion of the Phanerozoic time scale*). Subsequent investigators have determined that the designation Neogene correctly applies to the rock systems and corresponding time intervals delineated by Lyell, though some authorities prefer to exclude the Pleistocene from the Neogene. The Paleogene encompasses the Paleocene, Eocene, and Oligocene. (The terms Paleocene and Oligocene were coined subsequent to Lyell's work and inserted in the lower part of the Cenozoic stratigraphic scheme.)

Cenozoic rocks

Cenozoic rocks are extensively developed on all the continents, particularly on lowland plains, as, for example, the Gulf and Atlantic coastal plains of North America. They are generally less consolidated than older rocks, although some are indurated (cemented) as a result of high pressure due to deep burial, chemical diagenesis, or high temperature—namely, metamorphism. Sedimentary rocks predominate during the Cenozoic, and more than half the world's petroleum occurs in such rocks of this age. Igneous rocks are represented by extensive early Cenozoic flood basalts (those of East Greenland and the Deccan trap of India) and the late Cenozoic flood basalts of the Columbia River in Washington, as well as by numerous volcanoes in the circum-Pacific System and ocean island chains such as Hawaii.

Several of the world's great mountain ranges were built during the Cenozoic. The main Alpine orogeny, which produced the Alps and Carpathians in southern Europe and the Atlas Mountains in northwestern Africa, began roughly between 37 and 24 million years ago. The Himalayas were formed some time after the Indian Plate collided with the Eurasian Plate. These lofty mountains marked the culmination of the great uplift that occurred during the late Cenozoic when the Indian Plate drove many hundreds of kilometres into the underbelly of Asia. They are the product of the low-angle underthrusting of the northern edge of the Indian Plate under the southern edge of the Eurasian Plate.

From about five million years ago, the Rocky Mountains and adjoining areas were elevated by rapid uplift of the entire region without faulting. This upwarping sharply steepened stream gradients, enabling rivers to achieve greater erosional power. As a result, deep river valleys and canyons, such as the Grand Canyon of the Colorado River in northern Arizona, were cut into broad upwarps of sedimentary rock during late Cenozoic time.

On a global scale the Cenozoic witnessed the further dismemberment of the Northern Hemispheric supercontinent of Laurasia: Greenland and Scandinavia separated during the early Cenozoic about 55 million years ago and the Norwegian-Greenland Sea emerged, linking the North Atlantic and Arctic oceans. The Atlantic continued to expand while the Pacific experienced a net reduction in size as a result of continued seafloor spreading. The equatorially situated east–west Tethyan seaway linking the Atlantic and Pacific oceans was modified significantly in the east during the middle Eocene—about 45 million years ago—by the junction of India with Eurasia, and it was severed into two parts by the confluence of Africa, Arabia, and Eurasia during the early Miocene approximately 18 million years ago. The western part of the Tethys evolved into the Mediterranean Sea not long after it had been cut off from the global ocean system about 6 to 5 million years ago and had formed evaporite deposits which reach up to several kilometres in thickness in a land-locked basin that may have resembled Death Valley in present-day California. Antarctica remained centred on the South Pole throughout the Cenozoic, but the northern continents converged in a northward direction.

The global climate was much warmer during the early Cenozoic than it is today, and equatorial-to-polar thermal gradients were less than half of what they are at present. Cooling of the Earth began about 50 million years ago and, with fluctuations of varying amounts, has continued inexorably to the present interglacial climatic period. It is to be noted that a unique feature of the Cenozoic was the development of glaciation on the Antarctic continent about 35 million years ago and in the Northern Hemisphere between 3 and 2.5 million years ago. Glaciation left an extensive geologic record on the continents in the form of predominantly unconsolidated tills and glacial moraines, which in North America extend in a line as far south as Kansas, Illinois, Ohio, and Long Island, N.Y., and on the ocean floor in the form of ice-rafted detritus dropped from calving icebergs.

Extensive glaciation

Cenozoic life was strikingly different from that of the Mesozoic. The great diversity that characterizes modern-day flora is attributed to the explosive expansion and adaptive radiation of the angiosperms that began during the Late Cretaceous. As climatic differentiation increased over the course of the Cenozoic, flora became more and more provincial. Deciduous angiosperms, for instance, came to predominate in colder regions, whereas evergreen varieties prevailed in the subtropics and tropics.

Fauna also underwent dramatic changes during the Cenozoic. As was discussed in earlier sections, the end of the Cretaceous brought the eradication of dinosaurs on land and of large swimming reptiles (*e.g.,* ichthyosaurs, mosasaurs, and plesiosaurs) in marine environments. Nektonic ammonites, squidlike belemnites, sessile reef-building mollusks known as rudistids, and most microscopic plankton also died out at this time. The Cenozoic witnessed a rapid diversification of life-forms in the ecological niches left vacant by this great terminal Cretaceous extinction. In particular, mammals, which had existed for more than 100 million years before the advent of the Cenozoic Era, experienced substantial evolutionary radiation. Marsupials developed a diverse array of adaptive types in Australia and South America free from the predations of carnivorous placentals. The placental mammals, which today make up more than 95 percent of known mammals, radiated at a rapid rate. Ungulates (or hoofed mammals) with clawed feet evolved during the Paleocene (66.4 to about 57 million years ago). This epoch saw the development and proliferation of the earliest perissodactyls (odd-toed ungulates, such as horses, tapirs, rhinoceroses, and two extinct groups, the chalicotheres and titanotheres) and artiodactyls (even-toed ungulates, including pigs, peccaries, hippopotamuses, camels, llamas, chevrotains, deer, giraffes, sheep, goats, musk-oxen, antelopes, and cattle). During the later Cenozoic, perissodactyl diversity declined markedly, but artiodactyls continued to diversify. Elephants, which evolved in the late Eocene about 40 million years ago, spread throughout much of the world and underwent tremendous diversification at this time. Many placental forms of giant size, like the sabre-toothed cat, giant ground sloths, and woolly mammoths, inhabited the forests and the plains in the Pliocene (5.3 to 1.6 million years ago). It was also about this time that the first hominids appeared. Early modern humans, however, did not emerge until the Pleistocene.

Among marine life-forms, mollusks (primarily pelecypods and gastropods) became highly diversified, as did reef-building corals characteristic of the tropical belt. Planktonic foraminiferans underwent two major radiations—the first in the Paleocene and the second in the Miocene—punctuated by a long (15–20-million-year) mid-Cenozoic reduction in diversity related in all likelihood to global cooling.

Cenozoic life was affected significantly by a major extinction event that occurred between 10,000 and 8,000 years ago. This event, which involved the sudden disappearance of many Ice Age mammals, has been attributed to either of two factors: climatic change following the melting of the most recent Pleistocene glaciers or overkill by Paleolithic hunters. The latter is regarded by many as the more likely cause, as the rapidly improved technology of Paleolithic humans permitted more efficient hunting. (For a more detailed discussion of this matter, see below *Pleistocene Epoch: Pleistocene fauna and flora.*)

Late Cenozoic extinction event

TERTIARY PERIOD

General considerations. The Tertiary, the initial period of the Cenozoic Era, began about 66.4 million years ago and ended approximately 1.6 million years ago (see Table 4). The name Tertiary was introduced by Giovanni Arduino in 1760 as the youngest of a tripartite division of the Earth's rocks: the Primitive schists, granites, and basalts that formed the core of the high mountains (of Europe); the fossiliferous Secondary, or Mesozoic, in northern Italy (predominantly shales and limestones); and a younger group of fossiliferous sedimentary rocks, the Tertiary rocks, found chiefly at lower elevations. Although originally intended as a descriptive generalization of rock types, many of Arduino's contemporaries and successors gave these categories a temporal connotation and equated them with rocks formed prior to, during, and after the Noachian deluge.

In 1810 Alexandre Brongniart included all the sedimentary deposits of the Paris Basin in his *terrains tertiares,* or Tertiary, and soon thereafter all rocks younger than Mesozoic in western Europe were called Tertiary (Table 22). The recognition of the Quaternary Period in 1829 by Jules Desnoyers—based on the post-Tertiary deposits of the Seine valley—placed a somewhat different connotation on the term Tertiary, particularly in regard to its upper limits. Controversies regarding the connotation of the term Quaternary and its limits continue today in professional circles. Quaternary is not a satisfactory name in the hierarchy of stratigraphic nomenclature. The terms Primary and Secondary have been supplanted by Paleozoic and Mesozoic, and Tertiary is being gradually replaced by Paleogene and Neogene as formal period names in scientific literature (see below *Changing nomenclature*).

Subdivisions. The Tertiary faunas of western Europe that were known to 19th-century natural scientists consisted primarily of mollusks exhibiting varying degrees of similarity with modern types. At the same time, the science of stratigraphy was in its infancy, and the primary focus of its earliest practitioners was to use the newly discovered sequential progression of fossils in layered sedimentary rocks to establish a global sequence of temporally ordered stages of what was until that time an undivided record of Earth history. Lyell employed a simple statistical measure based on the relative percentages of living species of mollusks to fossil mollusks found in different layers of Tertiary rocks. These percentages had been compiled by his colleague and friend Gérard-Paul Deshayes, the French conchologist, who had amassed a collection of more than 40,000 mollusks and was preparing a monograph on the mollusks of the Paris Basin. In 1833, Lyell divided the Tertiary into four subdivisions (from older to younger): Eocene, Miocene, older Pliocene, and newer Pliocene (the latter was renamed the Pleistocene in 1839). The Eocene contained about 3 percent of the living mollusk species, the Miocene about 20 percent, the older Pliocene more than one-third and often over 50 percent, and the newer Pliocene about 90 percent. Lyell traveled extensively in Europe (and North America as well for that matter) and had a broad and comprehensive understanding of the regional geology for his day. He understood, for example, that rocks of the Tertiary were unevenly distributed over Europe and that there were no rocks of the younger part of the period in the Paris Basin. He used the deposits in the Paris and Hampshire and London basins as typical for the Eocene; the Faluns (shelly marls) of the Loire Basin near Touraine as well as the deposits in the Aquitaine Basin near Bordeaux in southwestern France and the Bormida River valley and Superga near Turin, Italy, for the Miocene; the sub-Apennine formations of northern

Lyell's subdivisions of the Tertiary

Table 22: Comparison of Some Cenozoic Classification Schemes

											European stages
Cenozoic	Pleistocene	original Pliocene of Lyell	Pleistocene	Pliocene	Pleistocene	Pleistocene	Pleistocene Plio-Pleistocene	Pleistocene *	Pleistocene Neogene	Neogene *	Calabrian
	Tertiary		Pliocenc		Pliocene	Pliocene	Pliocene	Pliocene *			Piacenzian
							Mio-Pliocene				Zanclean
		Miocene of Lyell	Miocene	Miocene		Miocene		Miocene *			Messinian
							Miocene				Tortonian
					Miocene						Serravallian
											Langhian
											Burdigalian
				Oligocene			Oligo-Miocene				Aquitanian
		Eocene of Lyell	Oligocene of Beyrich		Oligocene	Oligocene		Oligocene *	Paleogene	Paleogene *	Chattian
							Oligocene Eo-Oligocene				Rupelian
			Eocene	Eocene	Eocene	Eocene		Eocene *			Priabonian
							Eocene				Bartonian
			Paleocene of Schimper								Lutetian
							Paleocene-Eocene				Ypresian
					Paleocene	Paleocene *		Paleocene			Thanetian
							Paleocene				Selandian
							transition				Danian
Mesozoic	Cretaceous										

* Indicates recommended usage.

Italy for the older Pliocene; and the marine strata in the Gulf of Noto and the Island of Ischia (also in Italy) and Uddevalla (in Sweden) for the newer Pliocene.

The limits between Lyell's Tertiary subdivisions were not rigidly specified, and Lyell himself clearly recognized the approximate and imperfect nature of his scheme. Indeed, in their original form Lyell's subdivisions would today be termed biostratigraphic units (bodies of rocks characterized by particular fossil assemblages) rather than chronostratigraphic units (bodies of rocks deposited during a specific interval of time).

Subsequent stratigraphic studies in northern Europe showed that there were contemporaneous deposits that were included variously in the upper Eocene or lower Miocene by different geologists of the day. This situation led H.E. Beyrich, in 1854, to create the term Oligocene for rocks in the North German Basin and Mainz Basin and to insert it between the Eocene and the Miocene in the stratigraphic scheme. As originally proposed, the Oligocene included the Tongrian and Rupelian stages as well as strata that subsequently formed the basis for the Chattian Stage. The Tongrian is no longer used as a standard unit, its place being taken by the Rupelian. The term Paleocene was proposed by Wilhelm P. Schimper on the basis of fossil floras in the Paris Basin that he considered intermediate between Cretaceous and Eocene forms. Typical strata included the sands of Bracheux, the travertines of Sézanne, and the lignites and sandstones of Soissons (the Suessonian of Alcide D. d'Orbigny). As originally defined, the Paleocene was based on strata believed to be equivalent to the Ypresian Stage (the oldest unit of the Eocene), but it is now known in fact to lie stratigraphically below the oldest levels assigned to the Eocene in the Paris Basin. The problem of the Paleocene is that, of all the chronostratigraphic units of the Tertiary, it alone is defined on the basis of nonmarine strata, making recognition of its upper limit and general correlation difficult elsewhere. Acceptance of the term Paleocene into the general system of stratigraphic names was irregular, and only in 1939 did the United States Geological Survey, general arbiter of standard stratigraphic nomenclature in North America, formally accept it. The Danian Stage was proposed by the geologist Pierre Jean Édouard Desor in 1846 for chalk deposits in Denmark and was assigned to the Cretaceous by virtue of the similarity of its invertebrate megafossils to those of the youngest Cretaceous elsewhere. Since the late 1950s, micropaleontologists have recognized, however, that calcareous planktonic foraminiferans (pelagic, single-celled protozoans) and nannoplankton (marine phytoplankton) exhibit a major taxonomic change at the boundary between the Maastrichtian (uppermost Cretaceous) Stage and the Danian (lowermost Tertiary) Stage and that the affinities of both micro- and megafauna lie with the overlying Cenozoic. The Danian is now widely regarded as being the oldest stage of the Cenozoic.

In 1948 the 18th International Geological Congress placed the base of the Pleistocene at the base of the marine strata of the Calabrian Stage of southern Italy, using the initial appearance of northern or cool-water invertebrate faunas in Mediterranean marine strata as the marker. Subsequent studies showed that the type section was ill-chosen and that the base of the Calabrian Stage was equivalent to much younger levels within the Pleistocene. A newly designated stratotype section was chosen at Vrica in Calabria, and the base of the Pleistocene was found comparable to a level dated at nearly 1.65 million years. This was formally ratified by the 27th International Geological Congress in Moscow in 1984. The oldest stage of the Pleistocene, however, remains the Calabrian.

Changing nomenclature. As was noted above, a growing number of authorities consider the terms Tertiary and Quaternary to be outmoded. The Pleistocene is interpreted as equivalent to the Quaternary and the youngest of the epochs of the Neogene Period, whereas the Paleogene is regarded as the older part of the Cenozoic Era, equivalent to the Paleocene, Eocene, and Oligocene epochs. Inasmuch as substitution of the terms Paleogene and Neogene for Tertiary and Quaternary is still not universally accepted in the scientific community at this time, it seems appropriate to discuss Cenozoic history in terms of the Tertiary and the Quaternary. Nonetheless, the discussion of the controversies that currently exist is presented to demonstrate the fact that geology and, in particular, stratigraphy—like all sciences—is a dynamic and changing field, whose basic concepts are subject to modification as knowledge and understanding increase.

Tertiary rocks. *Types and distribution.* With the exception of the great Tethyan seaway, the basins of western Europe, and the extensive Mississippi embayment of the Gulf Coast region, Tertiary marine deposits are located predominantly along continental margins. They occur on all continents and are found in situ as far north as Alaska (Miocene), eastern Canada (Eocene), and Greenland (Paleocene). Deposits of Paleogene age occur on Seymour Island in the Antarctic Peninsula, and Neogene deposits containing marine diatoms (silica-bearing marine phytoplankton) have recently been identified intercalated between glacial tills on Antarctica itself.

Sedimentary deposits along continental margins

Global sea level is believed to have fallen gradually but inexorably about 300 metres over the past 100 million years, but superimposed upon that trend is a higher-order series of globally fluctuating increases and decreases in sea level with a periodicity of several million years. The resultant transgressions and regressions of the sea onto passive (*i.e.,* tectonically stable) continental margins has left a record of interfingered marine brackish and continental sedimentary deposits in Europe, North Africa, the Middle East, southern Australia, and the Gulf and Atlantic coastal plains of North America. In most regions, the Paleogene seas extended farther inland than did those of the Neogene; in fact, the most extensive transgression of the Tertiary is that of the Lutetian Age (Middle Eocene), about 49–45 million years ago, when the Tethys Sea expanded onto the continental margins of Africa and Eurasia and left extensive deposits of shallow-water carbonate rocks characterized by tropical foraminiferans of large size called *Nummulites* from Indonesia to Spain and as far north as Paris and London. Sediments of Tertiary age are widely developed on the deep ocean floor and elevated seamounts as well. In the shallower parts of the ocean above 4.5 kilometres, sediments are calcareous or siliceous (or both), depending on local productivity. Below 4.5 kilometres the sediments are principally siliceous or inorganic (*i.e.,* red clay) owing to dissolution of calcium carbonate.

Nonmarine (terrestrial, or continental, as they are called) Tertiary sedimentary and volcanic deposits are widespread in North America, particularly in the intermontane basins west of the Mississippi River. During the Neogene, volcanism and terrigenous deposition extended almost to the coast. In South America thick nonmarine clastic sequences (conglomerates, sandstones, and shales) occur in the mobile tectonic belt of the Andes Mountains and along their eastern front; these sequences extend eastward for a considerable distance into the Amazon Basin. Tertiary marine deposits occur along the eastern margins of Brazil and Argentina, and they were already known to Charles Darwin during his exploration of South America in 1833 and 1834.

Volcanism and orogenesis. Volcanism has continued throughout the Cenozoic on land and at the major oceanic ridges, such as the Mid-Atlantic Ridge and the East Pacific Rise, where new seafloor is continuously generated and carried away laterally by seafloor spreading. Iceland was formed in the middle Miocene, and it remains one of the few places where the processes that occur at the Mid-Atlantic Ridge can be observed today.

Massive outpourings of lava

Two of the most extensive volcanic outpourings recorded in the geologic record occurred during the Tertiary. The Deccan trap of India was the site of massive outpourings of basaltic lava near the boundary between the Cretaceous and Tertiary about 67–66 million years ago, whereas massive explosive volcanism took place near the Paleocene–Eocene boundary between 57 and 54 million years ago in northwestern Scotland, northern Ireland, and the Faeroe Islands and East Greenland, as well as along the rifted continental margins of both sides of the North Atlantic Ocean. This volcanic activity was associated with the initial rifting and separation of Eurasia and North America between Scandinavia and Greenland and left a stratigraphic record in the form of distal ash deposits as far south as the Bay of Biscay and the marine sedimentary basin of England. In both instances, comparable volumes of extensive basalts in the amount of 1 to 2 by 10,000,000 cubic kilometres were erupted. The well-known volcanics of the Massif Central of south central France, which figured so prominently in early (18th-century) investigations into the nature of igneous rocks, are of Oligocene age, as are those in central Germany. The East African Rift Zone preserves a record of mid-to-late Tertiary rifting and separation of the East African continent that eventually led to the formation of a marine seaway linking the Indian Ocean with the Mediterranean. During the Early Tertiary, volcanism occurred in the Caucasus Mountain region but was essentially absent elsewhere in Asia.

The circum-Pacific "Ring of Fire," an active tectonic belt that extends from the Philippines through Japan around the west coast of North and South America, was subject to seismic activity and andesitic volcanism throughout much of the Tertiary. The extensive Columbia Plateau basalts were extruded over Washington and Oregon during the Miocene, and many of the volcanoes of Alaska, Oregon, southern Idaho, and northeastern California date to the Late Tertiary. Active volcanism occurred in the newly uplifted Rocky Mountains during the Early Tertiary, whereas in the southern Rocky Mountains and Mexico volcanic activity was more common in the mid- and late Tertiary. The linear volcanic trends, such as the Hawaiian, Emperor, and Line island chains in the central and northwestern Pacific, are trails resulting from the movement of the Pacific Plate over volcanic "hot spots" (*i.e.,* magma-generating centres) that are probably fixed deep in the Earth's mantle. The major island groups such as the Hawaiian (which has been active over the past 30 million years), Galápagos, and Society islands (Miocene) are volcanoes that rose from the seafloor. Finally, Central America, the Caribbean region, and northern South America were the sites of active volcanism throughout the Cenozoic.

In contrast to the passive-margin sedimentation on the Atlantic and Gulf coastal plains, the Cordilleran (or Laramide) orogeny in the Late Cretaceous, Paleocene, and Eocene produced a series of upfolded and upthrusted mountains and deep intermontane basins in the area of the modern Rocky Mountains. Deeply downwarped basins accumulated as much as 8,000 metres of Paleocene and Eocene sediment in the Green River basin of southwestern Wyoming and 14,000 metres of sediment in the Uinta Basin of northeastern Utah. Other basins ranging from Montana to New Mexico accumulated similar but thinner packages of nonmarine fluvial and lacustrine sediments rich in fossil mammals and fish. In the Oligocene and Miocene, Cordilleran influences on what is now the western United States had ceased, and the basins were gradually filled to the top by sediments and abundant volcanic ash deposits from eruptions in present-day Colorado, Nevada, and Utah. These basins were exhumed during the Pliocene-Pleistocene with renewed uplift of the long-buried Rocky Mountains, along with uplift of the Colorado Plateau, producing steep stream gradients that resulted in the cutting of the Grand Canyon to a depth of more than 5,000 metres.

Volcanism along the Cascade Mountain chain was active from the late Eocene to today, as evidenced by the 1980 eruption of Mount St. Helens. This volcanism was gradually shut off in California as the movement of plate boundaries changed from one of subduction to a sliding and transform motion (see also EARTH, THE; *The major geologic features of the Earth's exterior: The surface of the Earth as a mosaic of plates; Types of plate boundaries*). With the development of the San Andreas Fault system, the western half of California started sliding northward. The Cascade–Sierra Nevada Mountain chain began to swing clockwise, causing the extension of the Basin and Range Province in Nevada, Arizona, and western Utah. This crustal extension broke the Basin and Range into a series of north–south-trending fault-block mountains and downdropped basins, which filled with thousands of metres of upper Cenozoic sediment. These fault zones (particularly the Wasatch Fault in central Utah and the San Andreas zone in California) remain active today and are the source of most of the damaging earthquakes in North America. The Andean mountains were uplifted during the Neogene as a result of subduction of the South Pacific beneath the South American continent.

Alpine orogeny

Complex tectonic activity occurred in Asia and Europe during the Tertiary. The main Alpine orogeny began during the late Eocene and Oligocene and continued throughout much of the Neogene. Major tectonic activity in the eastern North Atlantic (Bay of Biscay) extended into southern France and culminated in the uplift of the Pyrenees in the late Eocene. On the south side of the Tethys, the coastal Atlas Mountains of North Africa experienced major uplift during this time, but the Betic region of southern Spain and the Atlas region of northern Morocco continued to display mirror-image histories of tectonic activity well into the late Neogene. In the Middle East the suturing of Africa and Asia occurred about 18 million years ago. Elsewhere, India had collided with the Asian continent about 45 million years ago, initiating the Himalayan uplift that was to intensify in the late Neogene (*i.e.,* Pliocene and Pleistocene) and culminate in the uplift of the great Tibetan Plateau and the Himalayan Mountain range. Major orogenic movement also occurred in the Indonesian-Malaysian-Japanese arc system during the Neogene. In New Zealand, which sits astride the Indian-Australian and Pacific plate boundary, the major tectonic uplift (the Kaikoura orogeny) of the southern Alps began about 10 million years ago.

Sedimentary sequences. Northwestern Europe contains a number of Tertiary marine basins that essentially rim the North Sea Basin, itself the site of active subsidence during the Paleogene and basinal infilling during the Neogene. The marine Hampshire and London basins, the Paris Basin, the Anglo-Belgian Basin, and the North German Basin have become the standard for comparative studies of the Paleogene part of the Cenozoic, whereas the Mediterranean region (Italy) has become the standard for the Neogene. The Tertiary record of the Paris Basin is essentially restricted to the Paleogene (namely, those of Paleocene–late Oligocene age), whereas scattered Pliocene–Pleistocene deposits occur in England and Belgium above the Paleogene. The strata are relatively thin, nearly horizontal, and often highly fossiliferous, particularly in the middle Eocene *calcaire grossier* of the Paris Basin from which a molluscan fauna of more than 500 species has been described. The Paris Basin is a roughly oval-shaped basin centred on Paris, whereas the Hampshire and London basins lie to the southwest and northeast of London, respectively. The London Basin and the Anglo-Belgian Basin were part of a single sedimentary basin across what is now the English Channel during the early part of the Paleogene. The total Paleogene stratigraphic succession in these basins is less than 300 metres and consists of clays, marls, sands, carbonates, lignites, and gypsum, reflecting alternations of marine, brackish, lacustrine, and terrestrial environments of deposition. The alternating transgressions and regressions of the sea have left a complex sedimentary record punctuated by numerous unconformities and associated temporal hiatuses, and the correlation of these various units and events has challenged stratigraphers since the early 19th century. The integration of biostratigraphic zonal schemes based on calcareous marine protozoans, phytoplankton and armoured dinoflagellates, paleomagnetic stratigraphy, and tephrochronology (ash-bed correlations) has resulted in a refined correlation of the stratigraphic succession in these separate basins.

In contrast, extensive Tertiary sediments occur on the Atlantic and Gulf Coastal plains and extend around the margin of the Gulf of Mexico to the Yucatán Peninsula, a distance of more than 5,000 kilometres. Seaward these deposits can be traced from the Atlantic Coastal Plain to the continental margin and rise and in the Gulf Coastal Plain into the subsurface formations of this petroliferous province of the Gulf of Mexico. During the Paleocene the embayment of the Gulf Coastal Plain extended inland to southwestern North Dakota and Montana about 2,000 kilometres from the present shoreline, where the lower Paleocene (Danian) Cannonball Formation was deposited. Although eroded between northwestern South Dakota and southern Illinois, marine outcrops continue southward to the coastline where they continue in the subsurface of the Gulf of Mexico. Tertiary sediments with a thickness in excess of 15,000 metres are estimated to lie beneath the continental margin in the Gulf of Mexico. In the Tampico embayment of eastern Mexico, thicknesses of more than 3,000 metres have been estimated for the Paleocene Velasco Formation, which developed under conditions of active subsidence and associated rapid deposition. Exposures in the Atlantic Coastal Plain and most of the Gulf Coastal Plain are of Paleogene age, but considerable thicknesses of Neogene sediment occur in offshore wells in front of the Mississippi delta, where thicknesses in excess of 10,000 metres have been recorded for the Neogene alone. Sediments are dominantly calcareous in the Florida region and become more marly and eventually sandy and muddy to the west, reflecting the input of terrigenous matter transported seasonally by the Mississippi River and its precursor(s). A large number of local stratigraphic names are used, but outcropping Gulf Coast units of Paleogene age are now placed, in ascending order, in the Midwayan, Sabinian, Claibornian, Jacksonian, Vicksburgian, and Chickasawhayan stages. Overlying Neogene sediments, mostly known in the subsurface, are placed in the Anahuacan, Napoleonvillian, Duck Lakean, Clovellyan, and Foleyan stages. Owing to general faunal and floral similarities, it is possible to make relatively precise stratigraphic correlations in the Paleogene between the Gulf and Atlantic Coastal Plain region and the basins in northwestern Europe.

Tertiary sediments on the Atlantic and Gulf coastal plains

Correlation. The boundaries of the Tertiary were originally only qualitatively estimated on the basis of the percentages of living species of (primarily) mollusks in the succession of marine strata in the western European basins. Early correlations were made by direct correlations with the faunas in the type areas in Europe. It was soon realized that faunal provincialization led to spurious correlations, and in 1919 an independent set of percentages for the Indonesian region was proposed, which was subsequently modified into the so-called East India Letter Stage classification system based on the occurrence of taxa of larger foraminiferans. In this system, the Tertiary *a* corresponds to upper Paleocene, *Ta2* to lower Eocene, *Ta3* to middle Eocene, *Tb* to upper Eocene, *Tc* to lower Oligocene, *Td* to middle Oligocene, lower *Te* to upper Oligocene, upper *Te* to lower Miocene, lower *Tf* to middle Miocene, upper *Tf* to upper Miocene, *Tg* to lower Pliocene, and *Th* to upper Pliocene.

In Europe the need for more precise correlations of Mesozoic and Cenozoic marine strata led to the concept of stages, which was introduced in 1842 by d'Orbigny. These stages were originally defined as rock sequences composed of distinctive assemblages of fossils that were believed to change abruptly as a result of major transgressions and regressions of the sea. This methodology has since been improved and refined, but it forms the basis for modern biostratigraphic correlation.

Geochronology and microfossil zones. Since about the mid-1900s, increasing efforts have been made to apply radioisotopic dating techniques to the development of a geochronologic scale, particularly for the Cenozoic (see above *Relative and absolute dating: Absolute dating*). The decay of potassium-40 to argon-40 has proved very useful in this respect, and recent refinements in mass spectroscopy and the development of laser-fusion dating involving the decay of argon-40 to argon-39 has resulted in the ability to date volcanic mineral samples in amounts as small as single crystals with a margin of error of less than 1 percent over the span of the entire Cenozoic Era.

Also, since the mid-1960s, investigators have demonstrated that the Earth's magnetic dipole field has undergone numerous reversals in the past and that most rocks pick up and retain the magnetic orientation of the field at the time they are formed through either sedimentary or igneous processes. With the development of techniques for measuring the original magnetization, a sequence of polarity reversals has been dated for the late Neogene and a paleomagnetic chronology built up for the entire Cenozoic. This work is based on the recognition that the magnetic lineations detected on the ocean floor were formed when basaltic magma which had extruded from the oceanic ridges assumed the ambient magnetic polarity; the resulting strips of normal and reversed polarity reflected the magnetic reversals observed in deep-sea cores. Calibration of the composite geomagnetic polarity succession to time and the relation of this chronology to the isotopic time scale, however, have proved to be the greatest source of disagreement over various current versions of the geologic time scale. Calibrations of a time scale must ultimately be based on the application of meaningful isotopic ages to the succession of polarity intervals and geologic stages. A geochronologic scheme is thus an integration of several methodologies; it makes use of the best attributes of seafloor spreading history (*i.e.,* pattern of seafloor magnetic anomalies), magnetostratigraphy, and biostratigraphy in the application of relevant isotopic ages to derive a high-resolution and internally consistent time scale. The recent application of astronomically forced cyclical components of the stratigraphic record, such as lithological couplets of marl and chalks, fluctuations in the ratios and percentages of fossil taxa, and so forth, with a periodicity of 100,000 years, has resulted in fine-tuning the geologic time scale to a resolution of about 5,000 years in the late Neogene.

Over the past few decades, micropaleontologists have created a number of zones based on the regional distribution of calcareous plankton (foraminiferans and nannoplankton) and those of the siliceous variety (radiolarians and diatoms), making it possible to correlate sediments from the high northern to high southern latitudes by way of the equatorial region. The resulting high-resolution zonal biostratigraphy and its calibration to an integrated geochronology provide the framework in which a true historical geology has became feasible.

Boundaries and chronologies. Precise stratigraphic positions for the boundaries of the various Tertiary series were not specified by early workers in the 19th century. It is only in more recent times that the international geologic community, working mainly through the International Geological Congress and under the inspiration and leadership of Hollis D. Hedberg, has formulated a philosophical framework for stratigraphy by delineating what might be termed the holy trinity of stratigraphic concepts—lithostratigraphy, biostratigraphy, and chronostratigraphy. As has been already mentioned, by specifying the (lower) limits of rock units deposited during successive increments of geologic time at designated stratotype points in the rock record, geologists established a series of calibration points at which time and rock coincide. These boundary stratotypes are the linchpins of global chronostratigraphic units and serve as the point of departure for global correlation. In recent years the Eocene–Oligocene boundary has been stratotypically established in southern Italy, with a currently estimated age of 34 million years, and the Pliocene–Pleistocene boundary in Calabria, southern Italy, with a numerical age estimate of close to 1.65 million years. The Miocene–Pliocene boundary is stratotypified in Sicily and has been dated at about 5 million years ago, although the location of this boundary may be repositioned in the future. The Cretaceous–Cenozoic boundary has been stratotypified in Tunisia in North Africa; its estimated age is 66.4 million years. The Paleocene–Eocene boundary is currently under investigation and has an estimated age of 57–55 million years. The Oligocene–Miocene boundary (which corresponds to that between the Paleogene and Neogene) also is under study; its age has been calculated to be roughly 23.7 million years. An example of Neogene chronology and correlations is shown in Table 23.

Boundary stratotypes

Tertiary environment. *Paleogeography.* The present-day continent–ocean configuration is the result of a complex sequence of events involving the dynamic evolution and geometric rearrangement of the major landmasses and oceans that began almost 200 million years ago. By the beginning of the Cenozoic the continent–ocean geometry had assumed an essentially modern, or recent, aspect with several notable exceptions. The fragmentation and dispersal of the Southern Hemispheric supercontinent Gondwana continued in the Cenozoic. Australia separated from Antarctica in the late Paleocene (Figure 38), and the initial subsidence of the South Tasman Rise (at the eastern end of the Australia–Antarctica marginal contact) in the late Eocene resulted in a shallow but inexorably widening connection between the Indian and Pacific oceans (Table 24). The injection of relatively warm eastward-flowing currents and associated evaporation at relatively high latitudes set the stage for the initiation of glaciation on Antarctica by early Oligocene time about 34 million years ago. Progressive separation of the two continents led to the initiation of the circum-Antarctic Current, which sweeps around Antarctica and thermally isolates it from the effects of warmer waters and climates to the north.

The junction of India and Asia occurred during the middle Eocene approximately 45 million years ago and resulted in an effective, though not total, blockage of the westward-flowing Tethys. This was achieved about 18 million years ago with the junction of Africa and Asia near present-day Iran. Although the eastern and western Tethyan seaway was now severed, brief intermittent marine connections were reestablished 14 to 13 million years ago.

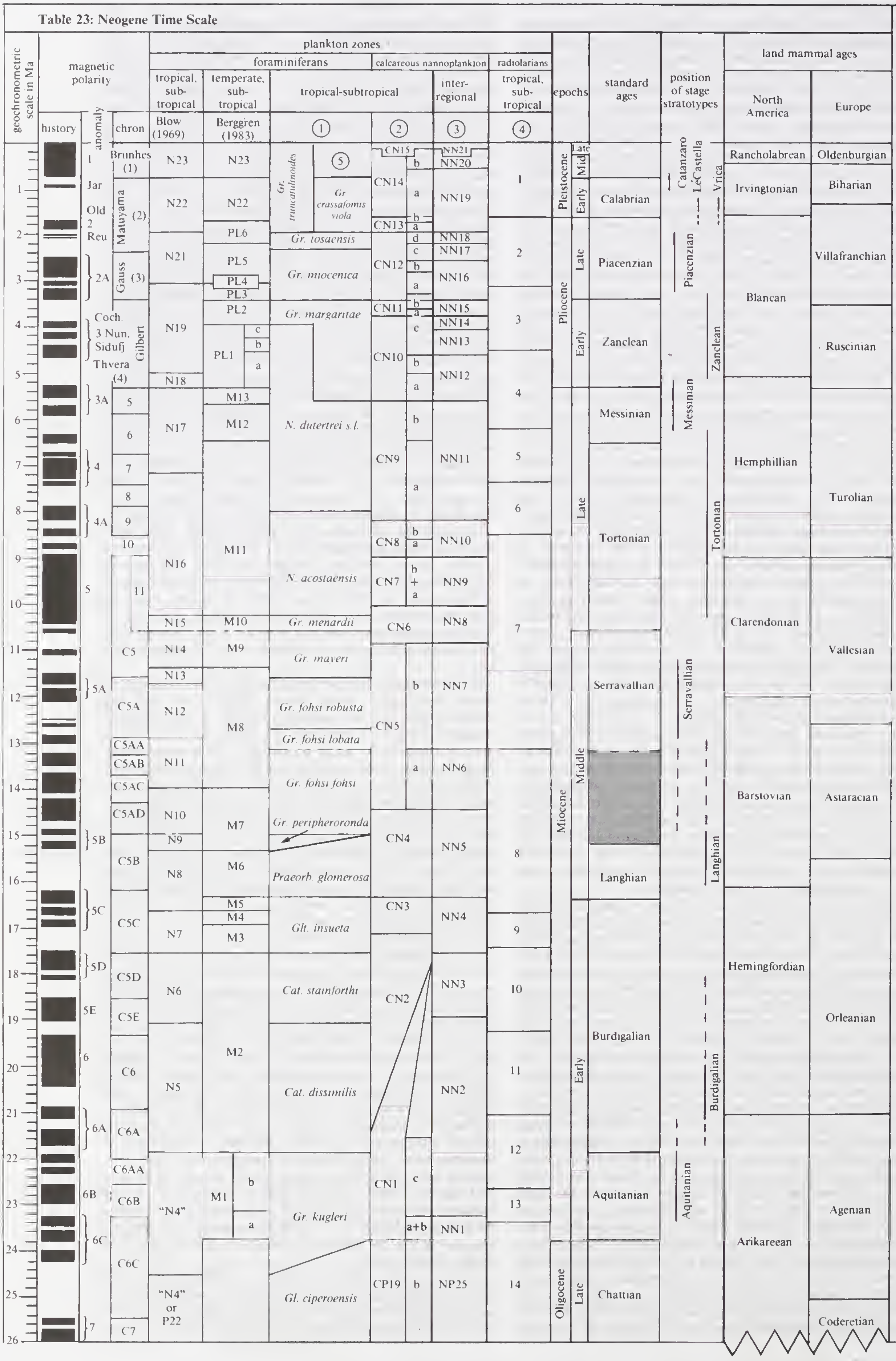
The present-day Mediterranean Sea is the geologically recent descendant of the Tethys. Between six and five million years ago the western remnant of the formerly extensive Tethyan seaway was subject to a brief (approximately one-million year) paroxysm that saw the entire basin virtually isolated from the world ocean; it experienced severe desiccation and the precipitation of a vast suite of evaporite deposits which reach up to several kilometres in thickness. The basin was subsequently refilled by the Atlantic and underwent significant geologic evolution during the past five million years. About one million years ago this part of the ancient Tethys was transformed into the Mediterranean Sea by the elevation of the Gibraltar sill and the consequent isolation of the basin from deep oceanic bottom waters and development of the present-day circulation pattern (Table 25).

Formation of the Mediterranean Sea

In the Northern Hemisphere the fragmentation and separation of Eurasia was completed during the early Paleogene with the opening of the Norwegian-Greenland Sea about 56 to 55 million years ago. Breaching of the subsiding subaerial Greenland-Scotland Ridge—formed during the Hebridean-Greenland eruptive volcanic episode of the late Paleocene mentioned above—allowed exchange of surface water between the Arctic and Atlantic oceans. Climatic conditions remained subtropical at high latitudes during the Paleogene as attested to by the remains of molluscan and shark faunas of tropical affinities in Spitsbergen. Furthermore, a fauna featuring such forms as the boid snake and durophagous alligator (a variety possessing teeth designed to crush food), as well as anguid and varanid lizards, emydid turtles, plagiomenids (flying lemurs), and paromomyids (primates), has been discovered on Ellesmere Island in the Canadian Arctic Archipelago, whose latitude has remained essentially stable—77° N—during the Cenozoic.

On the Eurasian continent the Ural Trough, a marine seaway linking the Tethys with the Arctic region that had constituted a barrier to the east–west migration of terrestrial faunas, was terminated by regional uplift at the end

Table 23: Neogene Time Scale



plankton zones ① After Bolli and Premoli-Silva (1973) Rögl and Bolli (1973); Stainforth et al. (1975) ② Bukry (1973, 1975) Okada and Bukry (1980) ③ Martini (1970, 1971) ④ Riedel and Sanfilippo (1978)

radiolarian zones
1. *Lamprocyrtis haysi*
2. *Pterocanium prismatium*
3. *Spongaster pentas*
4. *Stichocorys peregrina*
5. *Ommatartus penultimus*
6. *Ommatartus antepenultimus*
7. *Cannartus petterssoni*
8. *Dorcadospyris alata*
9. *Calocycletta costata*
10. *Stichocorys wolffii*
11. *Stichocorys delmontensis*
12. *Cyrtocapsella tetrapera*
13. *Lychnocanoma elongata*
14. *Dorcadospyris ateuchus*

Published with permission of the Geological Society of America

Table 24: Major Paleogene Paleogeographic Events*

age	paleogeographic events
Middle Oligocene (33–30 Ma)	Isolation of Antarctica completed after further subsidence of South Tasman Rise
Early Oligocene (*c.* 35–33 Ma)	Tethys severely restricted in the eastern part due to uplift of the Himalayas
Early Oligocene (37–35 Ma)	Separation of Greenland and Svalbard and availability of higher-latitude water to North Atlantic
Early Oligocene (38–35 Ma)	Shallow connection between the South Pacific and Atlantic developed at Drake Passage
Eocene/Oligocene Boundary (34–35 Ma)	Completion of the opening of Labrador Sea that had begun in the Maastrichtian
Late Eocene (*c.* 38 Ma)	Iceland–Faeroe sill sinks below sea level for the first time
Late Eocene (40–37 Ma)	Tethys partially restricted north and east of the Indian Plate
Late Eocene (*c.* 40 Ma)	Subsidence of South Tasman Rise permits shallow connection between Indian and Pacific oceans
Early Eocene (*c.* 53 Ma)	Opening Norwegian–Greenland Sea develops surface water exchange with the Arctic Ocean
Late Paleocene (55–53 Ma)	Australia and Antarctica separate and Australia begins northward drift; formation of the ocean between the two continents
Early Paleocene (56–55 Ma)	Separation of Greenland and Scandinavia and the formation of the Norwegian–Greenland Sea begins

*Listed are those paleogeographic events that affected global ocean circulation and certain climatic and faunal/floral migration patterns.

of the Eocene (Figure 39). The resulting immigration of Eurasian faunas into western Europe and the consequent faunal changes that occurred in terrestrial vertebrate faunas is known as the Grande Coupure (Big Break) among vertebrate paleontologists.

Relatively small changes in land–sea geometry have played an important role in the migration of terrestrial faunas and ultimately in the evolution of life itself. For example, during the early Paleogene land mammal exchange between Europe and North America occurred freely via a northern route owing to the close proximity of Spitsbergen, eastern Canada, and the subaerial Greenland-Scotland Ridge. The separation of the former two and the partial subsidence of the latter about 50 to 49 million years ago (middle Eocene) led to the termination of this free interchange and the development of separate evolutionary patterns among terrestrial vertebrate faunas in Europe and North America. The only route for faunal exchange between Eurasia and North America was the Bering Land Bridge that united Siberia and Alaska. It

Adapted from C R Scolese, The University of Texas at Arlington

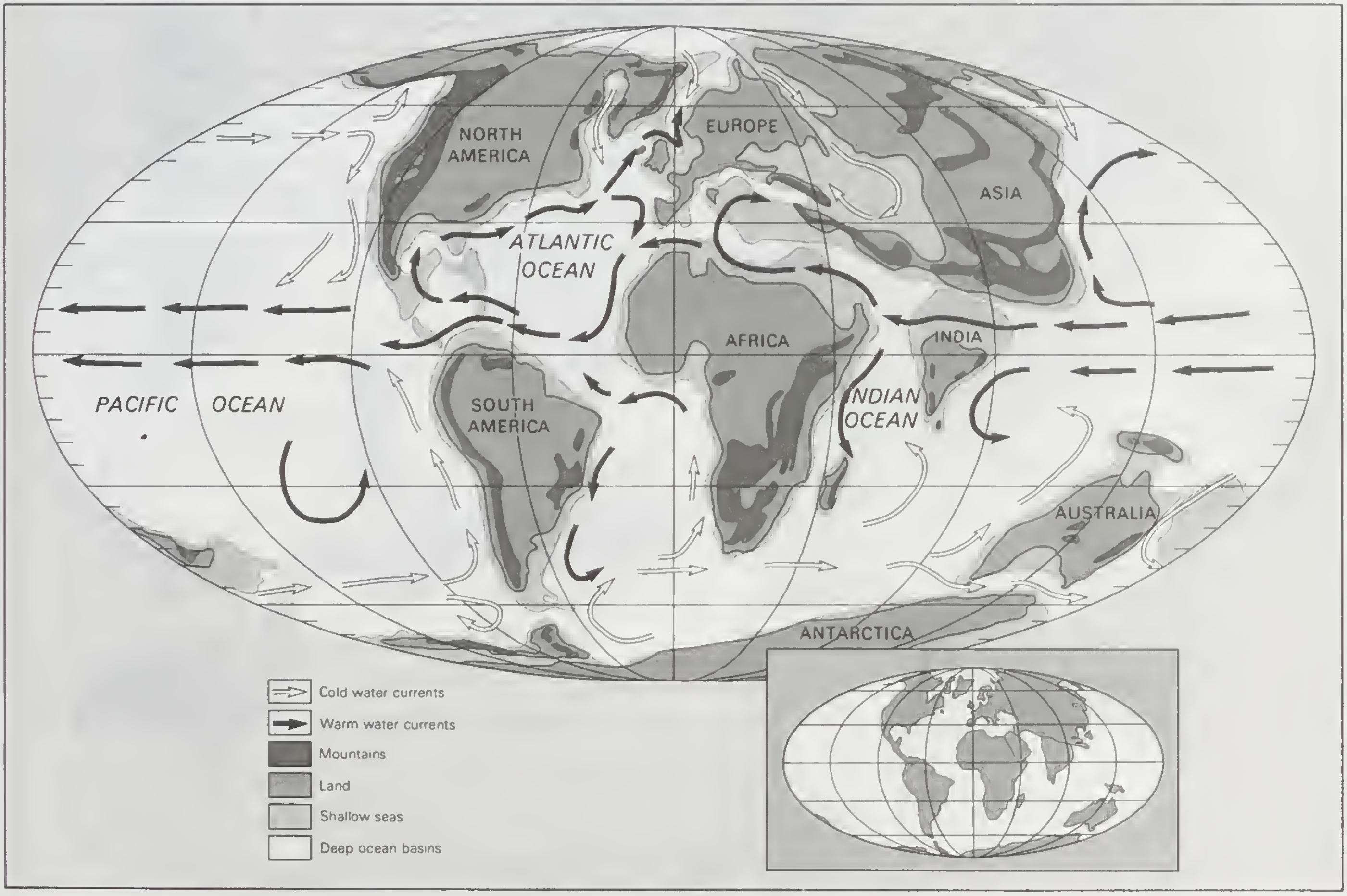


Figure 38: Distribution of landmasses, mountainous regions, shallow seas, and deep ocean basins during early Tertiary time. Included in the paleogeographic reconstruction are cold and warm ocean currents. The present-day coastlines and tectonic boundaries of the configured continents are shown in the inset at the lower right.

Table 25: Major Neogene Paleogeographic Events*

age	paleogeographic events
Pleistocene (*c.* 1 Ma)	Uplift of Gibraltar sill and development of present-day Mediterranean circulation patterns (surface water inflow, deep water outflow)
Middle Pliocene (3 Ma)	Uplift of Panamanic Isthmus, joining North and South America
Early Pliocene (*c.* 5 Ma)	Opening of Straits of Gibraltar
Late Miocene (*c.* 5.5 Ma)	Closure of Betic and Riffian (Moroccan) corridor, isolation of western Tethyan Sea from global ocean circulation, and evaporation of the basin
Middle Miocene (*c.* 13 Ma)	Final severance of the Tethys and Paratethys (epeiric continental seaway in southwestern Eurasia)
Early Miocene (*c.* 18 Ma)	Junction of Africa and Eurasia

*Listed are those paleogeographic events that affected global ocean circulation and certain climatic and faunal/floral migration patterns.

seems to have been breached only in the past 2.5 million years, allowing the transit of cold water currents from the Pacific into the Atlantic. Evidence for this occurs in the form of North Pacific cryophylic molluscan faunas in the mid-Pliocene faunas of Iceland. To the south, the Atlantic and Pacific oceans had been linked since the Early Cretaceous by the Panamanic Seaway in the Central American–northwest Colombian region. This seaway prevented terrestrial faunal interchange between North and South America, with the possible exception of a brief interlude during the Paleocene. It was closed by the elevation of the Isthmus of Panama about three million years ago with two significant geologic results. First, the emergence of the Isthmus of Panama permitted a major migration in land mammal faunas between North and South America—the so-called Great American Interchange—which saw South American ground sloths in North America in areas as dispersed as California, the Great Plains, and Florida, and North American faunas as far south as Patagonia. Second, the emergence of the Isthmus of Panama deflected the westward-flowing North Equatorial Current northward where it enhanced the northward-flowing Gulf Stream. The latter then carried warm, salty waters into high northern latitudes, contributing to greater precipitation through evaporation over the region of eastern Canada and Greenland and eventually to the development of the polar ice cap, which began forming between 3 and 2.5 million years ago in the Northern Hemisphere.

Emergence of the Isthmus of Panama and its effects

Significant geologic events. The concept of dynamic paleogeography provides a unifying framework within which to understand the causal link between changes in oceanic circulation, climate, and evolution—which together constitute geologic events. These events may be divided into two categories: physical and biotic.

The early Eocene opening of the Norwegian-Greenland Sea completed the fragmentation of the Northern Hemispheric supercontinent Laurasia and eventually united the Atlantic and Arctic oceans (Figure 40 and Table 24), although modern circulation patterns were not achieved until the subsidence of the Greenland–Scotland Ridge about 15 million years ago. In the Southern Hemisphere the separation of Australia and Antarctica reached a critical point about 34 million years ago, at which time the continent of Antarctica was covered by a major ice sheet. The junction of Eurasia and Africa about 18 million years ago severed the once more extensive Tethyan seaway, and the western part evolved, after being cut off from the world ocean for a relatively brief time, into the modern-day Mediterranean Sea (see above). Finally, the emergence of the Isthmus of Panama about 3 million years ago and concomitant changes in ocean circulation patterns led to the formation of a polar ice cap shortly thereafter in the late Pliocene. The history of the Earth over the past 2.5 million years has been intimately linked with repeated oscillations between glacial advances and retreats.

Biotic events reflect changes in paleogeography and climate. Among mammals the earliest equids (horses) and primates appeared during the early Eocene—a time of

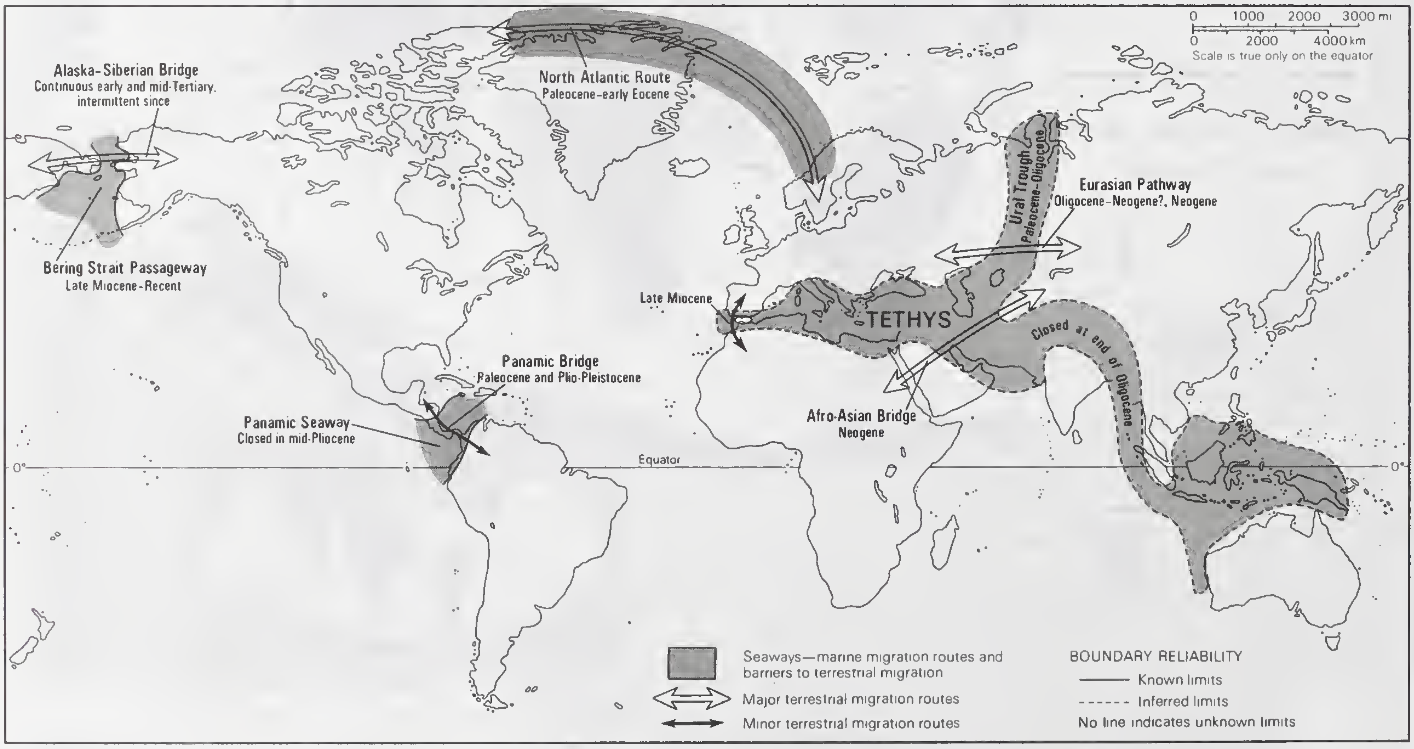

0 1000 2000 3000 mi
0 2000 4000 km
Scale is true only on the equator
Alaska-Siberian Bridge
Continuous early and mid-Tertiary, intermittent since
Bering Strait Passageway
Late Miocene-Recent
North Atlantic Route
Paleocene-early Eocene
Ural Trough
Paleocene-Oligocene
Eurasian Pathway
Oligocene-Neogene?, Neogene
Late Miocene
TETHYS
Closed at end of Oligocene
Panamic Bridge
Paleocene and Plio-Pleistocene
Panamic Seaway
Closed in mid-Pliocene
Afro-Asian Bridge
Neogene
Equator
0°
0°
▨ Seaways—marine migration routes and barriers to terrestrial migration
⇔ Major terrestrial migration routes
↔ Minor terrestrial migration routes
BOUNDARY RELIABILITY
Known limits
Inferred limits
No line indicates unknown limits

Figure 39: Principal Cenozoic faunal migration routes and barriers.

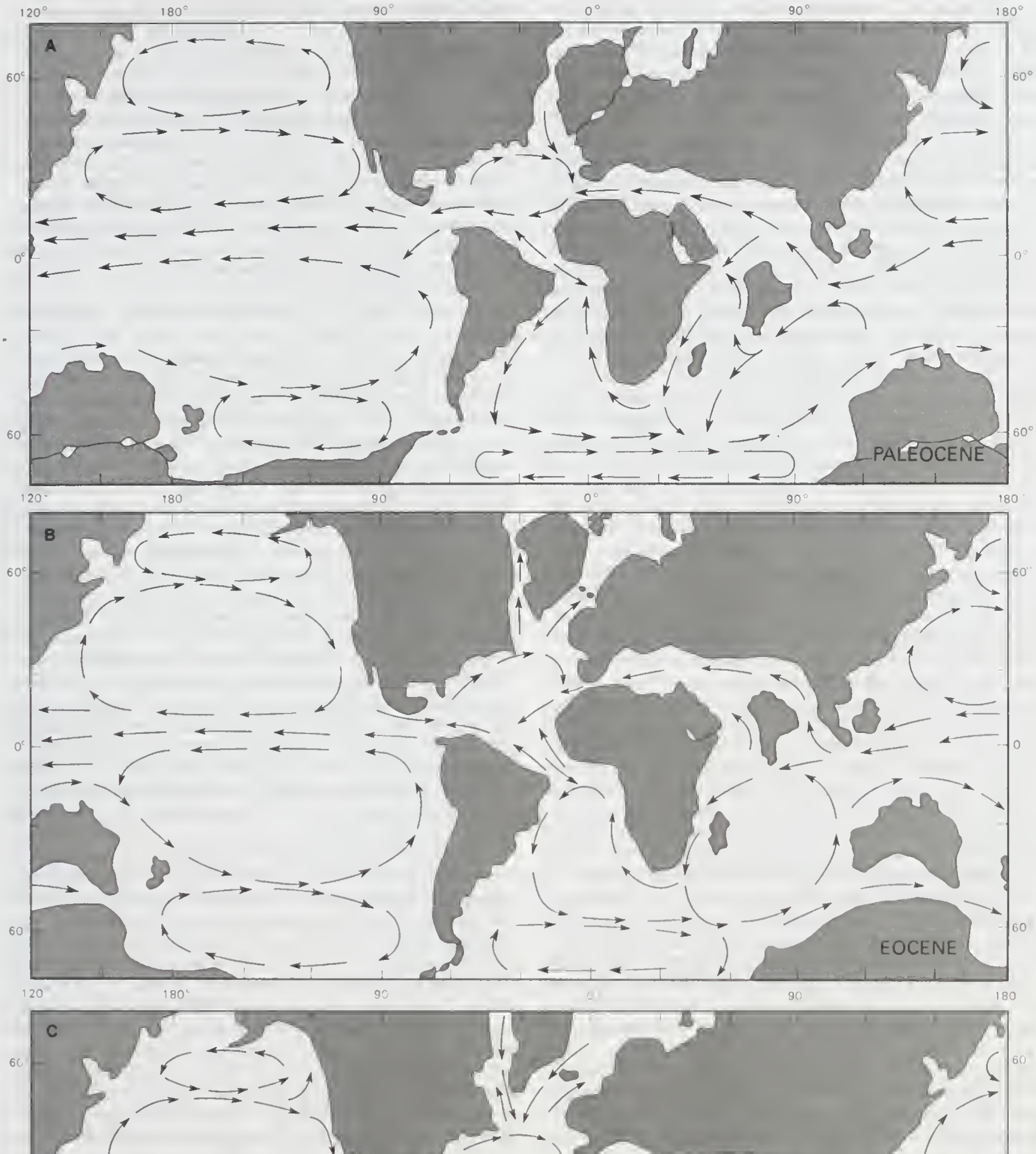


Figure 40: Inferred global ocean surface circulation patterns of the Paleogene plotted on a series of paleogeographic maps. (A) Clockwise-circulating subpolar gyres occur in the absence of a circum-Antarctic current system in the early Paleocene. (B) The Tethyan current north of India is restricted as the Indian Plate approaches the Eurasian continent during the mid-Eocene. Clockwise southern Atlantic and Pacific gyres persist owing to the disrupted flow of circulation around Antarctica resulting from continued closure of the Drake Passage. (C) During the late Oligocene, Tethyan flow is more restricted, having been diverted to the south of India after that continent's junction with Eurasia in the mid-Eocene. Cool water currents issue from the Arctic regions, and the circum-Antarctic current thermally isolates Antarctica from the world ocean, giving rise to colder climates and glacial conditions on Antarctica.

Migration of mammalian faunas

diversification of mammals. In the middle Eocene free land-mammal faunal migration between North America and Europe was interrupted by the severance of the land-bridge connection that had existed prior to this time. Although Europe was cut off from North America, Asia (especially Siberia) remained in contact with the latter in the late Eocene, and repeated migrations occurred throughout the Oligocene and Miocene. In the early Miocene, the first wave of immigration from Europe occurred, bringing bear-dogs, European rhinoceroses, weasels, and a variety of European deerlike animals to North America. In the early Miocene, mastodons escaped from their isolation in Africa, and by the middle Miocene they had reached North America. Rodents and early anthropoids evolved in the middle Eocene and the first elephants (Proboscidea) in the early Oligocene. Immigration of African mammalian faunas, including proboscideans, into Europe occurred about 18 million years ago (early Miocene).

The earliest apparent hominids have been reported in East Africa at about six to five million years ago and the subsequent australopithecine–hominid evolution in East Africa has been traced over the past three million years. In the late Pliocene, the Panamanic land bridge allowed porcupines, armadillos, and ground sloths to migrate from South America and live in the southern United States. A much larger wave of typically Northern Hemispheric animals, however, moved south and drove most of the South American endemic mammals to extinction. These North American invaders included dogs and wolves, raccoons, cats, horses, tapirs, llamas, peccaries, and even mastodons.

In the deep sea several major biotic events stand out. A major extinction event at the boundary between the Mesozoic and Cenozoic eras, 66.4 million years ago, affected dinosaurs, large marine reptiles, marine invertebrate faunas (rudists, belemnites, ammonites), and planktonic protozoans (foraminiferans) and phytoplankton (see above). On the other hand, deep-sea benthic protozoans suffered no effect until about 10 million years later at the boundary between the Paleocene and the Eocene, when more than half of all species suffered extinction under conditions that still remain unexplained but may be linked to changing deep-water circulation patterns. The present-day psychrospheric (*i.e.*, cold), benthic fauna evolved in the deep sea in the late Eocene about 36 to 35 million years ago, concomitant with significant cooling of oceanic deep waters of some 3°–5° C. The closure of the Tethyan seaway in the late early Miocene about 15 million years ago resulted in the disappearance of many of the larger tropical nummulitid-type foraminiferans that had ranged from Indonesia to Spain during most of the Tertiary. Although the descendants of these forms can be found today in the Indo-Pacific region, they show much less diversity.

The marine faunas of the eastern Pacific and West Indies–Caribbean region were similar throughout the Tertiary until about three million years ago. The elevation of the Isthmus of Panama at that time created a land barrier between the two regions that resulted in faunal provincialization.

Paleoclimate. Climatic history is intimately linked to the dynamic evolution of ocean-continent geometry and associated changes in oceanic circulation. The continued fragmentation of the world ocean due to changing positions of the main continental masses—principally a poleward shift in the Northern Hemisphere—led to increasingly inefficient latitudinal thermal-energy exchange. Paleobiogeographic and oxygen-isotope studies yield a complementary picture of a long-term global temperature decline, development of a thermally stratified ocean, and enhanced climatic differentiation during the Cenozoic. This climatic decline followed a faunally and florally recognizable climatic optimum in the early Eocene, which is also reflected in the oxygen-isotope records. In general terms Mesozoic oceanic circulation was latitudinal (and meridional transport of heat energy was relatively inefficient), whereas Cenozoic circulation has been predominantly longitudinal (meridional), although meridional heat transport has become increasingly less efficient during the Neogene as global temperatures have decreased.

Intensified climatic differentiation

During the early Paleogene, warm equable climates extended from pole to pole, with pole-to-equator temperature gradients of about 5° C during the Paleocene as compared to about 25° C today. The early Eocene witnessed the warmest conditions of the entire Cenozoic, with subtropical floras occurring on the margins of the Hampshire and London basins of southeastern England and varanid lizards, emydid turtles, alligators, and flying lemurs living on Ellesmere Island in the Canadian Arctic Archipelago. These circumstances attested to the presence of subtropical climates at 77° N latitude during this climatic optimum. Global cooling occurred during the middle and late Eocene and accelerated rapidly across the Eocene–Oligocene boundary at which time Antarctic continental glaciation was initiated.

Sea-level ice sheets had developed on West Antarctica during the early Oligocene and over most of the continent by the middle Miocene about 13 million years ago. Glaciation on the Antarctic continent in the late Miocene about 5.5 million years ago has been linked with the isolation of the Mediterranean Basin from the world ocean and its transformation into a desiccated basin not unlike present-day Death Valley for about 500,000 years. Mountain glaciers occurred in the Gulf of Alaska by the mid-Miocene and were followed by glaciers in Patagonian Argentina during the early Pliocene. The large ice sheets that covered northern Europe and North America first expanded about 3 million years ago, but major growth occurred 2.5 million years ago, at which time the Earth may be said to have passed over a thermal threshold initiating the so-called Ice Age, in which mode the Earth is still locked today. Repeated waxing and waning of the Northern Hemispheric glaciers over the past 2.5 million years resulted in significant and repeated expansions of the high-latitude belts of westerly winds toward the equator, changes in ocean circulation pattern (deflection of the Gulf Stream to an essentially east–west transit at about 40° N latitude), and the southward displacement of cool, dry climatic belts to southern Europe and North Africa during the cold periods.

As has been seen, the gradual breakup of Pangaea and Laurasia, closure of the Tethys Sea, and closure of the Panamanic Seaway have combined to provide greater provinciality in marine faunas during the Neogene.

Tertiary life. The end of the Mesozoic Era marked a major transition in Earth history. Major extinctions took place among marine and terrestrial animals; plant life suffered to a much lesser extent. The cause of this major event, whether single or multiple, is still being widely debated among specialists (see above *Mesozoic Era: Cretaceous Period: Cretaceous life: Mass extinctions*). In any case, the net result in the oceans was a marked reduction in diversity, primarily of calcium carbonate-secreting organisms (*i.e.*, coccolithophorids and planktonic foraminiferans), followed by a gradual recovery and radiation of new forms within a few hundred thousand years. The present-day ecosystem is for the most part populated by animals, plants, and single-celled organisms that survived and redeployed after the great extinction event at the end of the Mesozoic. Deep-sea benthic foraminiferans, mollusks, and teleost fishes survived and became prominent elements in the Paleogene seas. Following the extinction of the reef-building rudists at the end of the Cretaceous, reef-building corals recovered by the Eocene, and their low-latitude, continuous stratigraphic record is taken as an indicator of the persistence of the tropical realm. Whales (cetaceans) are descended from carnivorous terrestrial mesonychid Condylarthra; they had become adapted to the marine environment by the middle Eocene. Another enormous marine carnivore was the shark, which descended from an essentially similar form of the early Jurassic. Other new forms in the late Paleogene seas were the penguins, a group of swimming birds, and the pinnipeds, a group that includes seals, sea lions, and walruses.

New marine forms

The Cretaceous–Paleogene transition was not marked by any significant change in terrestrial floras. Angiosperms continued the radiation that had begun in the mid-Cretaceous about 100 million years ago. Grasses were present by late Paleocene time, but they did not expand to form the upland grasslands and prairies that are intimately linked to expansion of grazing animals until late Oligocene and Miocene time.

Evolution and distribution of foraminiferans. Since about the 1960s detailed studies have shown that the calcareous planktonic protozoan foraminiferans (superfamily Globigerinacea) have evolved rapidly and dispersed widely, following a major extinction at the end of the Cretaceous. These organisms have proved to be extremely useful in regional and global correlation of oceanic sediments and uplifted marine strata found on land. Differential rates of evolution within different groups give rise to the greater utility of some forms in stratigraphic zonation and correlation than others. For example, conical species of the Paleogene *Morozovella* and Neogene *Globorotalia* have stratigraphic ranges that vary from one to five million years.

The larger foraminiferans—the nummulitids—were a group of circular-to-elliptical, shallow-water, tropical, benthic forms that had complex, labyrinthine interiors and internal structural supports to strengthen their adaptation to life in high-energy environments. They contained symbiotic algae in life conditions and received nourishment from symbiosis performed by the entrained algae. The genus *Nummulites* occurred in massive numbers and large size (diameters up to 150 millimetres) during the great middle Eocene transgression and formed extensive limestone deposits in Egypt from which the pyramids were built. *Nummulites* lived throughout the Eurasian Tethyan province from late Paleocene to early Oligocene time but did not reach the New World. Following their extinction in the Oligocene, larger foraminiferans, the miogypsinids and lepidocyclinids, flourished during the Neogene. These forms are characterized by increased complexity in the internal part of the test (structural hard parts) by the addition of lateral chambers and changes in the geometry of the embryonic initial successive chambers. The miogypsinids ranged from the Oligocene to middle Miocene, while the lepidocyclinids disappeared in the early Pliocene, the last representative being recorded in Fiji.

Mammalian diversification and dominance

Vertebrates. In the terrestrial environment the most spectacular event of the Cenozoic has been the diversification and rise to dominance of the mammals. From only a few groups (opossums, archaic hoofed mammals, insectivorous mammals, and a number of extinct groups) that lived in the undergrowth hiding from the dinosaurs at the end of the Cretaceous, more than 20 orders of mammals evolved rapidly and were established by the early Eocene. During the Eocene, the first perissodactyls (such as primitive horses, rhinoceroses, and tapirs), artiodactyls (including camels and deer), rodents, and rabbits underwent wide dispersal, migrating via a northern Holarctic route, probably from Eurasia to North America. By the end of the Eocene, global climatic variations triggered changes in the vegetation from thick jungles to open forest/grasslands, causing extinction in most of the archaic browsing mammals typical of the Paleocene and Eocene. From the Oligocene onward, land mammal communities were dominated by groups living today, such as horses, rhinoceroses, antelopes, deer, camels, elephants, cats, and dogs. These groups, however, evolved significantly as the climate and vegetation changed to a more open, grassy habitat in the Miocene. Starting with primitive forms that had low-crowned teeth for browsing leafy vegetation, most of the herbivorous mammals evolved specialized teeth for grazing gritty grasses and long limbs for running and escaping more efficient predators. By the late Miocene, a savanna community analogous to that of the modern East African savanna was established on most continents. Beginning with the Messinian crisis of the late Miocene, climatic deterioration caused extinction in most of these mammals. Today, there remains but a pitiful remnant that has survived the subsequent Ice Age and the overhunting and habitat destruction caused by humans. (W.A.Be.)

QUATERNARY PERIOD

The Quaternary is both the shortest and most recent period of geologic time. It is the second period of the Cenozoic Era and began about 1.6 million years ago (see Table 4). The Quaternary is subdivided into two epochs, the Pleistocene and Holocene. The Pleistocene Epoch comprises almost all of Quaternary time; the Holocene, the latest and current interval, began a mere 10,000 years ago.

The term Quaternary originated early in the 19th century when it was applied to the youngest deposits in the Paris Basin in France. The designations Pleistocene and Holocene also date from the 19th century, when they were defined with respect to strata containing certain fossils or a certain percentage of fossils of plants and animals that were still living. Since their introduction, these terms have undergone a complex and confusing evolution with respect to their usage, and neither is used as originally defined.

In 1948 a decision was made at the 18th International Geological Congress in London that the Pliocene–Pleistocene boundary should be fixed in marine rocks exposed in the coastal areas of Calabria in southern Italy. As ratified by the International Commission on Stratigraphy in 1985, the type section for the Pliocene–Pleistocene boundary occurs in a sequence of marine strata at Vrica in Calabria (see below *Pleistocene Epoch: Stratigraphy*). A type section for the Pleistocene–Holocene boundary has yet to be agreed upon, but most investigators concur that it should be placed about 10,000 years ago.

Cycles of climatic change

The Quaternary is best characterized as a time of many cycles of climatic change. Some of these cycles resulted in the episodes of extensive glaciation of the Earth for which the Quaternary is well known. Although earlier studies suggested that there were four major glaciations, it is now recognized that many more occurred and that the climate of the Quaternary has alternated between periods of extensive ice buildup on land and periods of warmer climate, like today, when only about 10 percent of the land area is covered by glacial ice. The low-latitude regions of the Earth became increasingly arid during the Quaternary, and they were subject to fluctuating arid and more humid conditions. The climatic cycles and their resultant glaciations and periods of aridity had a dramatic effect on geologic processes, sedimentological regimes, the morphology of the terrestrial surface, and the fauna and flora on land and in the oceans. In addition, it was during the Quaternary that much of human evolution occurred, and these climatic cycles must have exerted a strong impact on the activities and distribution of early humans.

Understanding the Quaternary and its environments and climatic conditions is particularly important for interpreting past geologic time and for considering the future. According to the premise underlying the principles of uniformitarianism, studies of modern natural processes are the basis on which the past geologic record can be interpreted. Thus, studies of modern Holocene environments and processes provide the data base upon which inferences can be made on the origin and environments of older rocks, structures, landforms, and other Earth features. Of equal and probably greater importance are concerns regarding the current situation and the future. Present-day climatic conditions, ocean and continent configurations, and environments are similar to conditions that occurred during past Quaternary climatic cycles. Therefore, an understanding of the characteristics of these cycles and their cause provides a basis for predicting future climatic change and the environmental consequences of such change.

The significance of the Quaternary as a period of time was recognized more than a century ago. Grove K. Gilbert, an early prominent geologist of the United States Geological Survey, wrote in 1890: "When the work of the geologist is finished and his final comprehensive report written, the longest and most important chapter will be upon the latest and shortest of the geologic periods."

PLEISTOCENE EPOCH

The Pleistocene Epoch is best known as a time during which extensive ice sheets and other glaciers formed repeatedly on the landmasses and has been informally referred to as the "Great Ice Age." Modern research, however, has shown that large glaciers had formed prior to the Pleistocene—during the latter part of the Tertiary Period as well as during earlier periods of geologic time—and that glaciation is not unique to the Pleistocene.

Stratigraphy. *Pliocene–Pleistocene boundary.* Definition of the base of the Pleistocene has had a long and controversial history. Because the epoch is best recognized for glaciation and climatic change, many have suggested that its lower boundary should be based on climatic criteria—for example, the oldest glacial deposits or the first occurrence of a fossil of a cold-climate life-form in the sediment record. Other criteria that have been used to define the Pliocene–Pleistocene include the appearance of humans, the appearance of certain vertebrate fossils in Europe, and the appearance or extinction of certain microfossils in deep-sea sediments. These criteria continue to be considered locally, and some workers advocate a climatic boundary at about 2.4 million years.

Pre-Pleistocene intervals of time are defined on the basis of chronostratigraphic and geochronologic principles related to a marine sequence of strata. Following studies by a series of international working groups, correlation programs, and stratigraphic commissions, agreement was reached in 1985 to place the lower boundary of the Pleistocene series at the base of marine claystones that conformably overlie a specific marker bed in the Vrica section in Calabria. The boundary occurs near the level of several important marine biostratigraphic events and, more significantly, is just above the position of the magnetic reversal that marks the top of the Olduvai Normal Polarity Subzone, thus allowing worldwide correlation.

The Pleistocene is subdivided into informal time units, the early, middle, and late Pleistocene. The early Pleistocene extends to the Brunhes–Matuyama paleomagnetic boundary at 730,000 years ago, and the middle Pleistocene extends to the end of the next to the last glaciation at about 130,000 years. The late Pleistocene includes the last interglacial–glacial cycle ending at the Holocene boundary 10,000 years ago.

Chronology and correlation. The chronology of the Pleistocene originally developed through observation and study of the glacial succession, which in both Europe and the United States was found to contain either soils that developed under warm climatic conditions or marine deposits enclosed between glacial deposits. From these studies, as well as studies of river terraces in the Alps, a chronology was developed that suggested the Pleistocene consisted of four or five major glacial stages which were separated by interglacial stages with climates generally similar to those of today. Beginning with studies in the 1950s, a much better chronology and record of Pleistocene climatic events have evolved through analyses of deep-sea sediments, particularly from the oxygen isotope record of the shells of microorganisms that lived in the oceans.

Marine oxygen isotope record

The isotopic record is based on the ratio of two oxygen isotopes, oxygen-16 (^{16}O) and oxygen-18 (^{18}O), which is determined on calcium carbonate from shells of microfossils that accumulated year by year on the seafloor. The ratio depends on two factors, the temperature and the isotopic composition of the seawater from which the organism secreted its shell. Shells secreted from colder water contain more oxygen-18 relative to oxygen-16 than do shells secreted from warmer water. The isotopic composition of the oceans has proved to be related to the storage of water in large ice sheets on land. Because molecules of oxygen-18 evaporate less readily and condense more readily, an air mass with oceanic water vapour becomes depleted in the heavier isotope (oxygen-18) as the air mass is cooled and loses water by precipitation. When moisture condenses and falls as snow, its isotopic composition is also dependent on the temperature of the air. Snow falling on a large ice sheet becomes isotopically lighter (*i.e.,* has less oxygen-18) as one goes higher on the glacier surface where it is both colder and farther from the moisture source. As a result, large ice sheets store water that is relatively light (has more oxygen-16), and so during a major glaciation the ocean waters become relatively heavier (contain more oxygen-18) than during interglacial times when there is less global ice. Accordingly, the shells of marine organisms that formed during a glaciation contain more oxygen-18 than those that formed during an interglaciation. Although the exact relationship is not known, about 70 percent of the isotopic change in shell carbonate is the result of changes in the isotopic composition of seawater. Because the latter is directly related to the volume of ice on land, the marine oxygen isotope record is primarily a record of past glaciations on the continents (see Figure 41).

Long core samples taken in portions of the ocean where sedimentation rates were high and generally continuous and where water temperature changes were relatively small have revealed a long record of oxygen isotope changes that indicate repeated glaciations and interglaciations going back to the Pliocene. The record is relatively consistent from one core sample to the next and can be correlated throughout the oceans. Warmer periods (interglacials) are assigned odd numbers with the current warm interval, the Holocene, being 1, while the colder glacial periods are assigned even numbers. Subdivisions within isotopic stages are delineated by letters (see Figure 41). The ages of the stage boundaries cannot be measured directly, but they can be estimated from available radiometric ages of the cores and from position with respect to both paleomagnetic boundaries and biostratigraphic markers, and also by using sedimentation rates relative to these data.

The record for the last 730,000 years indicates that eight major glacial and interglacial events or climatic cycles of about 100,000 years' duration occurred during this interval. Similar cycles but of greater frequency and lesser magnitude continue back to the late Pliocene (see Figure 41). An isotopic record from the North Atlantic suggests the first major glaciation in that region occurred about 2,400,000 years ago, some 800,000 years before the start of the Pleistocene. It also suggests that the first glaciation likely to have covered extensive areas of North America and Eurasia occurred about 850,000 years ago during oxygen isotope stage 22. The largest glaciations appear to have taken place during stages 2, 6, 12, and 16; the interglacials with the least global ice, and thus possibly the warmest, appear to be stages 1, 5, 9, and 11. The last interglaciation occurred during all of stage 5 or just substage 5e, depending on location; the last glaciation took place during stages 4, 3, and 2; and the current interglaciation falls during stage 1.

The marine isotopic record is a continuous record, unlike most terrestrial records, which contain gaps because of erosion or lack of sedimentation and soil formation or a combination of these factors. Because of its continuity and its excellent record of climatic events on land (glaciations), the marine oxygen isotope record is the standard to which the terrestrial and other stratigraphic records are correlated. Correlations to it are based on available chronometric ages, on paleomagnetic data where available, and on attempts to match the terrestrial record and its interpretation with specific characteristics of the isotopic curve. Unfortunately, most terrestrial records contain few radiometric ages and are incomplete, and specific correlations, except for the most recent part of the record, are difficult and uncertain. A few terrestrial records, however, are exceptional and can be correlated with confidence.

Loess–paleosol records

Central China is covered by deposits of windblown dust and silt, called loess. Locally the loess is more than 100 metres thick, mantling hillsides and forming loess plateaus and tablelands. The loess accumulated primarily during times that were colder and drier than present, and most of it was derived from desert areas to the west. The loess succession contains many colourful buried soils or paleosols that formed during periods which were both warmer and wetter than today. Thus, on stable tablelands with minimal erosion, the succession provides an exceptional climatic and chronological record that extends back 2.4 million years to the late Pliocene. In total, up to 44 climatic cycles have been delineated, with more frequent cycles occurring during the early Pleistocene. Although not directly related to glaciation, correlation with the marine oxygen isotope record is excellent, and many of the specific loess and soil units have similar climatic inferences, as do their correlative oxygen-18 stages.

Another loess–paleosol succession occurs in the Czech Republic, Slovakia, and Austria, where loess blankets terraces of the major rivers that drained eastward and southward from the principal glaciated areas in the Alps and northern Europe. As in China, buried soils are common in the loess succession and, along with gastropod shells, provide

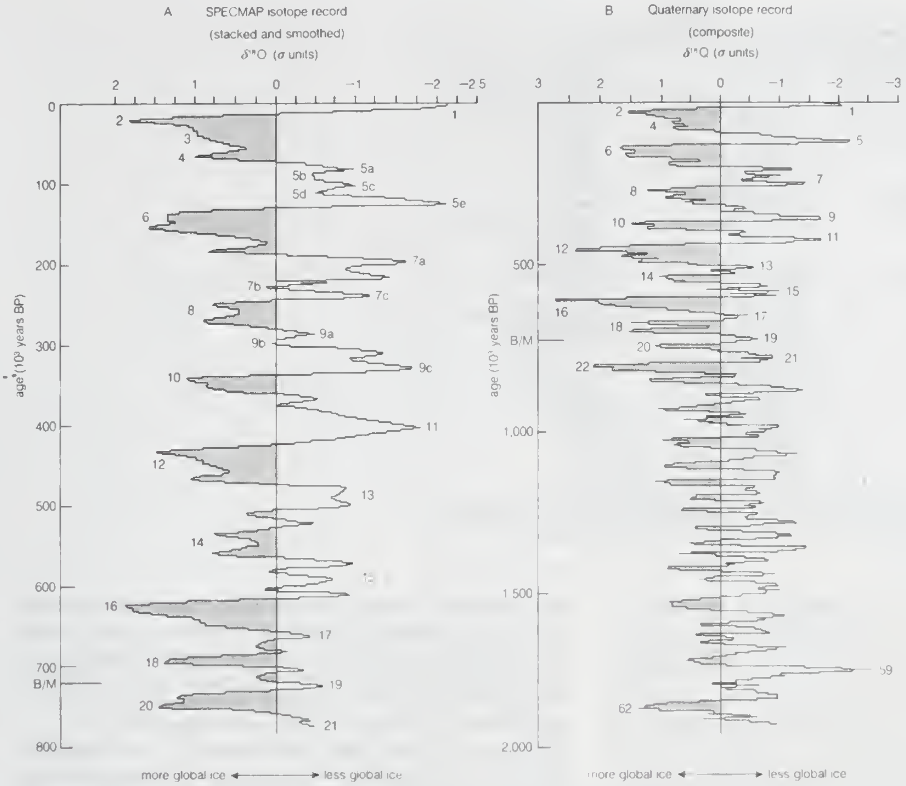


Figure 41: *Summary of marine oxygen isotope records.*
(A) The SPECMAP (Spectral Mapping Project) record based on five low- and middle-latitude deep-sea cores and (B) a composite record of four cores from the equatorial Pacific, the Caribbean, and the North Atlantic. Isotopic stages and substages are indicated; B/M shows the level of Brunhes/Matuyama reversal.

From N J Shackelton and N G Pisias (1985), D G Martinson *et al* (1987), J Imbrie *et al* (1984), and D F Williams *et al* (1988) in S C Porter *Quaternary Research*, 32, © 1989 University of Washington

paleoclimatic data and evidence for climatic change. The climatic cycles varied from cold and dry conditions when loess accumulated to warm and wet conditions with hardwood forests and well-developed soils. In the last 730,000 years, eight climatic cycles have been delineated; these correlate with the eight oxygen-18 cycles that occurred in the marine record during the same time interval. During the entire Pleistocene, about 17 glacial episodes alternated with 17 interglacials.

Glacial records

Glacial till, which was directly deposited by glaciers, covers extensive areas of northern Eurasia and northern North America and occurs as well in many mountain regions and other areas that currently are not covered by glacial ice. Soils of warm climate origin buried between tills were recognized long ago and provided the basis for the development of the idea of multiple glaciation during the Pleistocene. However, because direct dating of the deposits generally is not possible and the glacial sequence is not complete as a result of erosion or nondeposition or a combination of the two, the development of long chronological records and correlation to the oxygen-18 record are difficult. Correlations generally are possible for the last two climatic cycles. They also are feasible in areas where the glacial succession contains interbedded volcanic rocks from which radiometric ages can be obtained.

In the mid-continental region of the United States, early work recognized tills that were interpreted to represent four major glaciations and three major buried soils that were viewed as representing interglaciations (see Table 26). Subsequent work showed that the glaciated record was more complex and that parts of the older record were miscorrelated. Consequently, the older portion of the record is informally referred to as the pre-Illinoian, and the older glacial and interglacial terms are no longer used except locally. Volcanic ash occurs within the succession in Iowa, Kansas, and Nebraska and is useful for correlation and dating. In one core, till occurs below ash that has been dated at about 2.2 million years old, suggesting late Pliocene glaciation. Other tills of the pre-Illinoian sequence probably are correlative with oxygen-18 stages 22, 16, and 12, and possibly others. The Illinoian correlates with oxygen-18 stage 6 and possibly stage 8, and the Sangamonian correlates with stage 5. The last glacial interval, the Wisconsinan, is subdivided into three parts, an early stade (substage) of glaciation, a middle interstadial, or time of restricted glaciation, and a late stade of glaciation. These intervals generally correlate with oxygen-18 stages 4, 3, and 2, respectively. Deposits of the early and middle Wisconsinan are poorly known in the mid-continental region of the United States; the area probably was not glaciated. Tills of the early Wisconsinan and even some that are correlative with oxygen-18 substages 5d or 5b, however, are common in the Canadian Arctic and on Baffin Island, where the ice sheet developed much earlier. It was not until the late Wisconsinan, about 18,000 years ago, that the southern ice sheet margin reached its maximum extent in the United States and eastern and western Canada. The ice sheet margin began to retreat and downwaste (*i.e.,* thin out) soon after reaching its maximum position, and the United States was deglaciated by about 10,000 years ago. Hudson Bay, near the centre of the ice sheet, was open to the ocean by 8,000 years ago, and, except for the Barnes and Penny ice caps on Baffin Island, the ice sheet had dissipated from the upland areas of central Canada by 6,000 years ago, well into the Holocene and oxygen-18 stage 1.

A somewhat similar chronology has been developed for the glaciated areas of Eurasia and the British Isles based on a variety of criteria. In addition to tills and buried soils, marine deposits, permafrost features, and fossil pollen and beetles have been used to subdivide the succession on a climatic basis. As elsewhere, the earlier portion of the record is not well established, and correlations among different geographic areas, as well as to the marine oxygen-18 record, are uncertain (see Table 26). The first cold period, known as the Pretiglian and based on pollen data from The Netherlands, began about 2.3 million years ago, soon after e:tensive ice-rafted material first appears in North

Table 26: Classic Glacial/Cold and Interglacial/Warm Episodes*

oxygen-18 stage	central United States	Great Britain	northwestern Europe
1	**Holocene**	**Holocene, Flandrian**	**Holocene, Flandrian**
	Wisconsinan	Devensian	Weichselian
2	late	late	late
3	middle	middle	middle
4 or 5a–d	early	early	early
5 or 5e	**Sangamonian**	**Ipswichian**	**Eemian**
6, 8?	Illinoian	Wolstonian	Saalian
6			Warthe
8			Drenthe
	Yarmouthian†	**Hoxnian**	**Holsteinian**
12	Kansan†	Anglian	Elsterian
	Aftonian†	**Cromerian**	**Cromerian complex**
	Nebraskan†	Beestonian	**Bavel complex**
		Pastonian	
		Pre-Pastonian	Menapian
		Bramertonian	**Waalian**
		Baventian	Eburonian
		Antian	**Tiglian**
		Thurnian	
		Ludhamian	
		Pre-Ludhamian?	Pretiglian

*Interglacial/warm episodes in boldface; correlations between areas are not well established and are not intended for the early portion of the record. †Included informally in the Pre-Illinoian.

Atlantic deep-sea cores. The Pretiglian was followed by a succession of warm and cold intervals, which also are based on pollen and on other flora and fauna evidence and which have been given different names in different areas. Although several old gravels with glacial erratics are known, the oldest major glacial episodes with extensive till deposits are the Elsterian in northern Germany and the Anglian in England. These glaciations probably are correlative with oxygen-18 stage 12, and local evidence suggests the possibility of earlier glacial events. Along coastal areas, these tills are overlain by the marine Holstein deposits, which also may represent more than one high sea-level stand. The next major glacial sequence is the Saalian of Germany, which is subdivided into the Drenthe and the Warthe; these probably correlate with oxygen-18 stages 8 and 6, respectively. Deposits and soils of the last interglaciation, the Eemian and Ipswichian, are correlative with oxygen-18 stage 5e, and those of the last glaciation, the Weichselian and Devensian, correlate with oxygen-18 stages 5d–a, 4, 3, and 2. As in central North America, tills and other deposits are well known only from the last part of this interval. The deglacial history generally is similar, except for a widespread but short interval of renewed glacial activity and cold climatic conditions that is known as the Younger Dryas in Scandinavia and Loch Lomond in the British Isles. This event occurred about 11,000 years ago, some 2,000 years before the dissipation of the ice sheet.

Ice-core records

A relatively short but important late Pleistocene and Holocene climatic record is derived from ice cores that have been taken from the ice sheets of Antarctica, Greenland, and Arctic Canada. The ice record in several cores extends back to the last interglaciation (oxygen-18 stage 5) and, in one case, to the next-to-the-last glaciation (stage 6). Although dating of the lower portions of the ice cores is difficult, annual layers of snow and ice can be counted in the upper parts and an accurate time scale reconstructed. Because the air temperature at the time when moisture condenses to fall as snow controls the oxygen and hydrogen isotopic composition of the snow, investigators are able to reconstruct temperature variations through isotopic studies of the ice cores. Data from the Vostok core taken from the East Antarctic Ice Sheet indicate that the climatic record of the Southern Hemisphere is similar to that interpreted from Northern Hemisphere records with respect to times of glaciation and interglaciation (see Figure 42). It also is possible to measure the amount of microparticles (very fine dust) in the ice, and studies of this kind show that there are many more particles in the portions of the core that accumulated during periods of extensive glaciation, apparently reflecting greater atmospheric circulation and dust in the atmosphere at those times. Trapped air preserved in small bubbles in the ice gives an indication of the composition of the atmosphere at the time the ice (snow) accumulated. An important result from this work indicates that the amount of carbon dioxide in the atmosphere during the last glacial (stages 2, 3, and 4) was substantially less than during the Holocene (stage 1) and the last interglaciation (stage 5e). This observation has significant implications with respect to climate and climatic change during glacial and interglacial transitions.

Pleistocene events and environments. Environments during the Pleistocene were dynamic and underwent dramatic change in response to cycles of climatic change and the development of large ice sheets. Essentially all regions of the Earth were influenced by these climatic events, but the magnitude and direction of environmental change varied from place to place. The best-known are those that occurred from the time of the last interglaciation, about 125,000 years ago, to the present.

Glaciation. The growth of large ice sheets, ice caps, and long valley glaciers was among the most significant events of the Pleistocene. During times of extensive glaciation, more than 45 million square kilometres (or about 30 percent) of the Earth's land area were covered by glaciers, and portions of the northern oceans were either frozen over or had extensive ice shelves. In addition to the Antarctic and Greenland ice sheets, most of the glacial ice was located in the Northern Hemisphere, where large ice sheets extended to mid-latitude regions. The largest was the Laurentide Ice Sheet in North America, which at times stretched from the Canadian Rocky Mountains on the west to Nova Scotia and Newfoundland on the east and from southern Illinois on the south to the Canadian Arctic on the north (see Figure 43). The other major ice sheet in North America was the Cordilleran Ice Sheet, which formed in the mountainous region from western Alaska to northern Washington. Glaciers and ice caps were more widespread in other mountainous areas of the western United States, Mexico, Central America, and Alaska, as well as on the islands of Arctic Canada where an ice sheet has been postulated.

Although smaller in size, the Scandinavian Ice Sheet was similar to the Laurentide in character (see Figure 43). At times, it covered most of Great Britain, where it incorporated several small British ice caps, and extended south across central Germany and Poland and then northeast across the northern Russian Plain to the Arctic Ocean. To the east in northern Siberia and on the Arctic Shelf of Eurasia, a number of small ice caps and domes developed in highland areas, and some of them may have coalesced to form ice sheets on the shallow shelf areas of the Arctic Ocean. Glaciers and small ice caps formed in the Alps and in the other high mountains of Europe and Asia. In the Southern Hemisphere, the Patagonia Ice Cap developed in the southern Andes, and ice caps and larger valley glaciers formed in the central and northern Andes. Glaciers also developed in New Zealand and on the higher mountains of Africa and Tasmania, including some located on the equator.

Glaciers and ice caps in the Andes

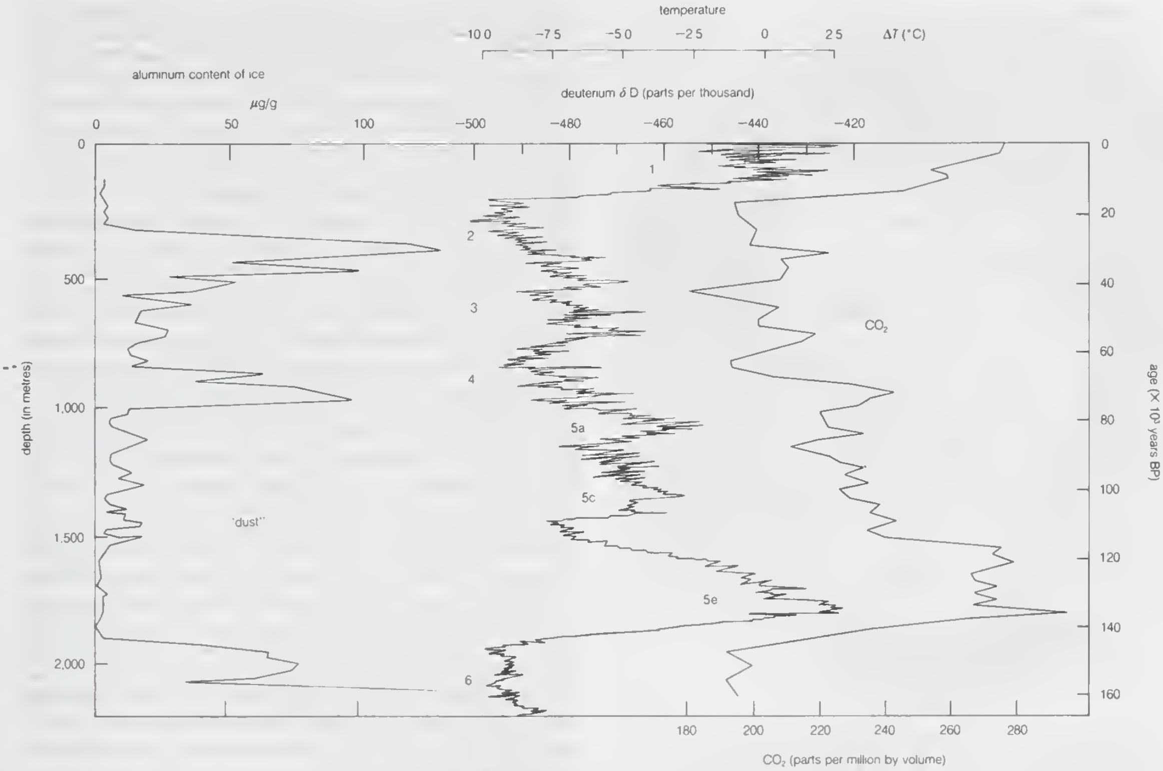


Figure 42: Dust content (as indicated by aluminum), temperature relative to current surface temperature (as indicated by deuterium), and CO_2 content of the Vostock ice core from the East Antarctic Ice Sheet. Correlative marine oxygen isotope stages are indicated.

From J Jouzel *et al* (1987) and M DeAngelis *et al* (1987) in W S Broecker and G H Denton, *Geochimica et Cosmochimica Acta*, vol 33, © 1989 Pergamon Press

The results of glaciation varied greatly, depending on regional and local conditions. Glacial processes were concentrated near the base of the glacier and in the marginal zone. Material eroded at the base was transported toward the margin, where it was deposited both at the glacier bed and in the marginal area. These processes resulted in the stripping of large quantities of material from the central zones of the ice sheet and the deposition of this material in the marginal zone and beyond the ice sheet. The Laurentide and Scandinavian ice sheets scoured and eroded bedrock terrain in their central areas, leaving behind many lakes and relatively thin glacial drift. On the other hand, the Central Lowland and the northern Great Plains of the United States and the western plains of Canada, as well as northern Germany and Poland, southern Sweden, and portions of eastern and northern Russia, contain relatively thick deposits of till and other glacial sediment. The landscape of such areas is flat to gently rolling. Today, these areas are among the great agricultural regions of the world, which is in large part attributable to glaciation.

The effects in mountainous terrain were even more dramatic. Glacial processes were concentrated in the upper regions where snow accumulated and in the valleys through which the glaciers moved to lower elevations. These valley glaciers carved towering peaks (such as the Matterhorn in the Alps), large rock basins, and sweeping U-shaped valleys and left some of the most spectacular scenery on the Earth, with many high-level lakes and waterfalls. The lower portions of the valleys commonly contain ridges of glacial drift. Ridges of this sort that form along valley slopes are called lateral moraines, while those that loop across a valley at the lower end of a glacier are termed end moraines. The earliest observations and interpretations of more extensive Pleistocene glaciation were made on such deposits and landforms in the Alps during the early part of the 19th century.

Periglacial environments. The environment around the ice sheets was markedly different from that of today in these formerly glaciated areas. Temperatures were much lower, and a zone of permafrost (perennially frozen ground) developed around the southern margin of the ice sheets in both North America and Eurasia. This zone was relatively narrow in central North America, on the order of 200 kilometres, but in Europe and Russia it extended many hundreds of kilometres south of the ice margin. Mean annual temperatures near the ice margin were about −6° C or colder and increased away from the ice margin to about 0° C near the southern extent of the permafrost. Compared to present-day conditions, the mean air temperature was on the order of 12° to 20° C colder near the ice margin. These conditions are indicated by ice-wedge casts and large-scale patterned ground, which are relict forms of ice wedges and tundra polygons that form today only in areas with continuous permafrost. Frost activity through freezing and thawing was intensified, and in areas of more relief talus accumulations and large block fields formed along escarpments and valley sides. Mass-wasting processes also were intensified and much material was eroded from slopes in periglacial areas. Deposits and landforms from such activity are known from the British Isles, northern Europe, and what was formerly the Soviet Union.

Lacustrine environments. Large lakes, usually many times bigger than their modern counterparts, were common during the Pleistocene. They fluctuated in level in response to the major climatic cycles or the opening and closing of outlets due to glaciation and vertical movements of land areas. Some lakes were closely tied to glaciation. In North America a series of large proglacial lakes formed around the margin of the Laurentide Ice Sheet during backwasting (recession) of the ice margin into Hudson Bay. The lakes were confined in part by the ice margin and in part by higher land to the south, east, and west. One of the largest was Lake Agassiz, which covered sizable areas of Manitoba, Ontario, and Saskatchewan and extended into North Dakota and Minnesota. The Great Lakes also formed as a result of glaciation as lobes of ice moved down preexisting lowlands and scoured out the weak rocks in the basins. Other lakes formed in the Champlain and Hudson valleys in eastern North America during deglaciation. Similar glacial lakes developed around the Scandinavian Ice Sheet and in other glaciated regions.

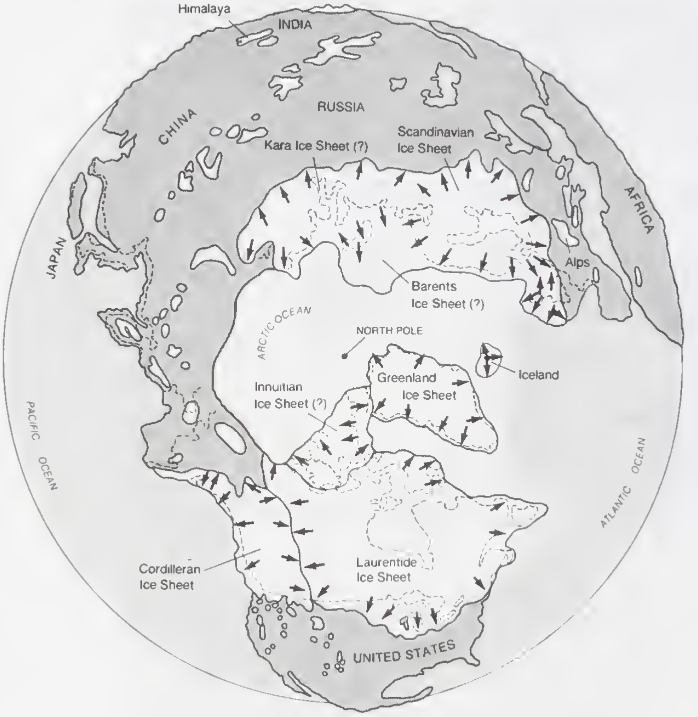


Figure 43: Areas of the Northern Hemisphere that were covered by glacial ice during the last glaciation. The arrows indicate the general direction of ice flow, and sea ice is shown covering the Arctic Ocean and extending south into the North Atlantic. The existence of the Barents, Kara, and Innuitian ice sheets remains open to dispute.

From B J Skinner and S C Porter, *Physical Geology* (1987), John Wiley & Sons, Inc

Pluvial lakes

Of equal interest was the development of large lakes in areas that today have arid to semiarid climatic regimes and generally lack lakes or have modern lakes that are much reduced in size and are saline in character. Such lakes are referred to as pluvial lakes, and the climate under which they existed is termed a pluvial climate. Most of these lakes existed in closed basins that lacked outlets, and thus their levels were related to relative amounts of precipitation and evaporation. A record of fluctuating lake levels is provided by ancient shorelines and beach deposits that are present along the slopes of the enclosing mountains as well as by the sediment and soil record preserved in the subsurface deposits of the lake basins. The history of lake fluctuations varies somewhat locally within a region but may be much different from one region of the world to another, depending on the local and regional climate.

In the Great Basin of Utah, Nevada, California, and Oregon and in other areas of the western and southwestern United States and Mexico, about 100 basins contained lakes during the Pleistocene (see Figure 44). The largest of these was Lake Bonneville, the predecessor of the modern Great Salt Lake in Utah. At its highest stage Lake Bonneville covered an area of about 52,000 square kilometres, and its maximum depth was approximately 370 metres. These conditions existed about 15,000 years ago during the interval of the last major Pleistocene glaciation. Lake Bonneville shrank rapidly in size and, by 12,000 years ago, had permanently shrunk to a point where it had become smaller than the Great Salt Lake. A long record of fluctuating lake levels is evident from a 930-metre core taken in the Searles Lake basin in California. Parts of the sediment record from the core sample indicate a deep lake with lacustrine silts and clays and freshwater fossils. Other parts contain unusual evaporite minerals which indicate that the lake was shallow and highly saline or even evidence of sediment exposure indicative of the complete dessiccation of the lake. The inferred climatic record from the core is similar to the marine oxygen isotope record but differs in that it shows more variation in the amplitude of the climatic cycles.

Pluvial lakes in these areas were most extensive during times of widespread glaciation in the Northern Hemisphere and were low or dry during times of reduced glacial cover. Paleoclimatic modeling suggests that the Laurentide Ice Sheet forced the polar jet stream south of its present-day position during glaciation. This brought more moisture from the Pacific into the desert areas of the southwestern United States, causing greater precipitation as well as producing more cloud cover, which, together with lower temperatures, resulted in less evaporation.

Pluvial lakes also were common in other dry regions of the world, particularly in the subtropical zones, including eastern and northern Africa and portions of Australia, Asia, and the Middle East. Examples of these pluvial bodies are the Dead Sea in Jordan and Israel and Lake Chad in the southern Sahara. The latter, now a shallow saline lake, covered some 300,000 square kilometres and was about six times the size of Lake Bonneville. A number of lakes in the rift valleys of East Africa were larger and deeper than they are today. Among the better-known and better-understood are Lakes Rudolf, Victoria, Nakuru, Naivasha, Magadi, and Rukwa. Most of these lakes in the tropical and subtropical regions were not in phase with those in the Great Basin of North America. They were relatively high for some 20,000 or more years immediately before the last glaciation and again just after the last glaciation in the early Holocene. A long climatic record inferred from sediments in Lake George in southeastern Australia has characteristics similar to those of the marine oxygen isotope record. Alternating humid and arid climatic cycles were more rhythmic and of greater magnitude in the middle and late Pleistocene than earlier, and a major change in basin hydrology occurred approximately 2.5 million years ago.

Rivers and river valleys

Fluvial environments. Rivers and the valleys that they occupy were affected strongly by the changing climates of the Pleistocene. River channels and their sediment record are controlled in large part by the amount and type of load that is supplied by their drainage basins and the discharge or quantity of water available for flow. Both are closely related to climate, which not only includes precipitation, evaporation, and seasonality but also controls the extent of

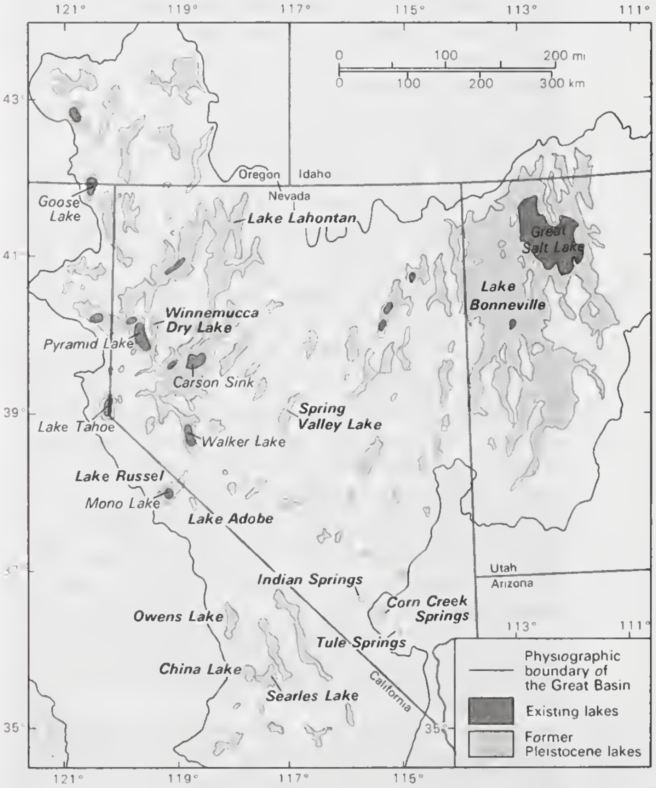


Figure 44: Distribution of late Pleistocene pluvial lakes and existing lakes in the Great Basin of the western United States.

Adaptation ot map "Lakes and Marshes ot Assumed Late-Pleistocene Age in the Great Basin," "Late Wisconsin Paleoecology of the American Southwest" by W G Spaulding, Leopold and Van Deuender *Late-Quaternary Environments of the United States*, H E Wright, Jr , ed , vol 1, *The Late Pleistocene*, Stephen C Porter, ed the vegetative cover of the land and the type and intensity of weathering processes. In addition, because of sea-level changes related to glaciation, the base level of rivers in coastal regions also fluctuated by significant amounts. As a result, river environments were dynamic and variable.

This was true for most rivers, but particularly so for those rivers that drained large quantities of meltwater and sediment from the glacier margins. During glaciation, rivers of the latter kind developed braided-channel patterns in response to the input of large quantities of sediment derived from the melting glaciers and subglacial waters and to the large fluctuations in the quantity of water flowing at any one time, which varied because of seasonal and diurnal controls on the generation of meltwater. During times of glaciation many of these rivers deposited thick sequences of sand and gravel in their valleys; examples include those of the Hudson, Mississippi, and Ohio rivers in the United States and of the Thames, Elbe, Rhine, and Seine rivers in Europe. Similar valleys have been buried by younger glacial deposits and are no longer evident at the surface. They exist today as bedrock valleys with thick fills of fluvial sand and gravel or lacustrine silt in localities where lakes existed in the valleys as a result of glacial damming. The sand and gravel fill in the surface valleys provide aggregate material for construction, and much groundwater is derived from the fills of both surface and buried valleys.

Some glacial valleys, as well as large upland areas, were sites of major catastrophic floods that resulted from the sudden drainage of proglacial and subglacial lakes. Such floods are known as jökulhlaups, an Icelandic term for subglacial lake outbursts. The largest and best-known floods of this type occurred in the Channeled Scabland of the Columbia Plateau region in eastern Washington state. Ice tongues flowing south from the Cordilleran Ice Sheet periodically dammed the Clark Fork River, forming glacial Lake Missoula. At times, Lake Missoula stretched more than 200 kilometres upvalley and was about 600 metres deep near the ice dam. Sudden failure of the ice dam released over 2,000 cubic kilometres of water, which flooded westward and southward across the Columbia Plateau and down the Columbia River valley. The floods cut through a loess cover into basalt and left a system of large dry channels with waterfalls, potholes, and longitudinal grooves in the basalt. Associated with the dry channels are huge, coarse gravel bars and giant current ripples. Other large catastrophic floods resulted from the sudden drainage of glacial Lake Agassiz and from the ancestral Great Lakes, as well as from some nonglacial lakes such as Lake Bonneville in the Great Basin (see above). During the Anglian–Elsterian glaciation in Europe a large ice-dammed lake formed in the North Sea, and large overflows from it initiated cutting of the Dover Straits.

During the transition from glacial to interglacial conditions, river channel patterns evolved from braided to meandering as a result of decreased load and possibly discharge. Near glaciated areas, rivers eroded into glacial oûtwash and left a system of stream terraces along the sides of most valleys. These modern interglacial rivers are much smaller than their glacial counterparts and are underfit (*i.e.*, appear too small) with respect to the large valleys in which they flow. In contrast, near coastal areas rivers actively built up their channels during the transition to interglacial conditions in response to rising sea level.

Coastal environments and sea-level changes. Coastal environments during the Pleistocene were controlled in large part by the fluctuating level of the sea as well as by local tectonic and environmental conditions. As a result of the many glaciations on land and the subsequent release of meltwater during interglacial times, sea level has fluctuated almost continuously between interglacial levels, like those of today, and levels during times of maximum glaciation, such as 18,000 years ago when sea level was more than 100 metres lower. At that time all the continental land areas were larger, and extensive areas of the world's continental shelves were exposed to weathering, soil formation, and fluvial and eolian activity and were inhabited by plants and animals. The Bering Shelf was exposed at this time and Siberia was connected to Alaska by a land bridge, thus allowing intercontinental migration of animals, including early humans. Rapid melting of the last large ice sheets resulted in a rising sea level that reached near modern level by the mid-Holocene, about 5,000 years ago. As a consequence, Pleistocene coastal environments are submerged below sea level in most parts of the world and are poorly known.

Fortunately some coastal areas of the world were undergoing tectonic uplift during the Pleistocene, and as a result older shorelines and their deposits are exposed above modern sea level. Study of these deposits is important in understanding the recent sea-level record and in relating it to the record of glaciation. The most important are shorelines that contain coral reefs, because it is possible to obtain radiometric ages on fossils in the reef complex. Two of the most important and best-dated records are on the island of Barbados in the Caribbean and along the Huron Peninsula of New Guinea (see Figure 45). The latter area exposes a spectacular suite of coastal terraces due to steady and rapid uplift during the Pleistocene. Age determinations of the terraces indicate times of relatively high sea level and suggest that they occurred at intervals of about 20,000 years. The highest sea level prior to the modern level occurred about 125,000 years ago and correlates with the peak warm interval of the last interglaciation (oxygen-18 stage 5e). Sea level at that time was about six metres higher than it is today.

From Steinan *et al* (1973) and J M A Chappell *et al* (1973) in A L Bloom *et al*, *Quaternary Research*, 4 © 1974 University of Washington

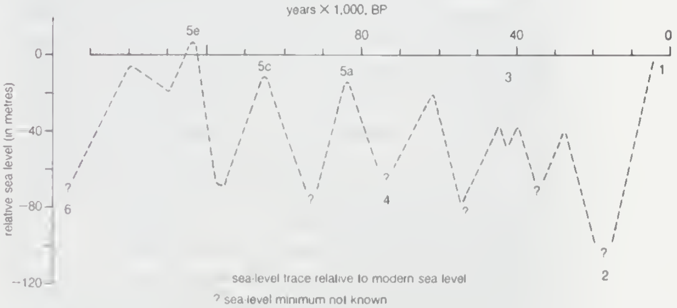


Figure 45: Late Pleistocene paleosea levels based primarily on data from New Guinea and Barbados. Correlative marine oxygen isotope stages are indicated.

Eolian activity. Eolian deposits are important in the Pleistocene record and indicate widespread wind action at certain times and in certain areas of the world. Mention has already been made of the importance of loess–paleosol records in working out regional chronologies and paleoclimatic history. Loess blankets large portions of the central and northwestern United States, Alaska, the east European plain of Russia, and southern Europe, where it is closely related to episodes of glaciation or to the cold periglacial climate beyond the ice sheet margins or to both. The loess was derived primarily from the broad floodplains of the braided rivers draining meltwater and sediment away from the glaciers as well as from newly exposed glacial drift. Locally, sand dunes and sheets of sand occur near the valley sources and in some cases cover large upland areas, as in central and northern Europe. The loess in China, on the other hand, is considered to have been deflated mostly from such desert areas as the Gobi.

Correlation between periods of peak aridity and peak glaciation

The deserts of the subtropical regions also experienced eolian activity during the Pleistocene. In Australia, the time of peak aridity and maximum dune activity (about 20,000 to 12,000 years ago) correlates with the time of peak glaciation in the Northern Hemisphere. This also was the case in the Sahara and other deserts in Africa, India, and the Middle East. One estimate is that the tropical arid zones were five times larger during times of peak glaciation. Sea level was lower at these times, the water was colder, and tropical cyclones were less extensive, resulting in decreased rainfall. These episodes of intensified eolian activity are recorded in other Pleistocene records. Ocean cores taken downwind of these regions contain windblown sediment in the portions of the core that accumulated during times of maximum eolian activity. In addition, microparticles occur in ice cores taken from the Greenland and Antarctic ice sheets and are concentrated at times of maximum glaciation and aridity in the subtropical deserts (see above Figure 42). At other times, the climate was less arid and the desert areas contracted, and vegetation developed to stabilize the dunes under more humid (pluvial) conditions.

Tectonic and isostatic movements. The lithospheric plates continued to shift during the Pleistocene, but the continents essentially were in their modern position at the start of the epoch. Of more importance to subsequent Quaternary events were the late Tertiary tectonic movements that affected the evolution of climate toward that of the Quaternary. Among these were the formation of the Isthmus of Panama, which affected oceanic circulation, and the uplift of the Tibetan Plateau and broad regional areas of the western United States, which affected atmospheric circulation, particularly the position and configuration of the polar jet stream.

Vertical movements of the Earth's crust also were caused by the formation and melting of large ice sheets. The area beneath an ice sheet subsides during glaciation because the crust is not able to sustain the weight of the glacier. These isostatic movements take place through the flow of material in the Earth's mantle, and the amount of subsidence amounts to about one-third the thickness of the ice sheet—for example, about one kilometre in the central area of the Laurentide Ice Sheet in Canada. Melting of the ice sheet removes the load and causes the ground to rise, or rebound. Such uplift is rapid at first but decreases with time. More than 300 metres of uplift has occurred in the eastern Hudson Bay area since that area was deglaciated. Substantial uplifting also took place prior to the complete melting of the ice sheets, and upward crustal movement continues today at a maximum rate of about 1.3 centimetres per year. A similar record of glacio-isostatic adjustments is encountered in Fennoscandia, where the greatest depression and subsequent uplift related to the Scandinavian Ice Sheet is located in the Gulf of Bothnia.

Pleistocene fauna and flora. The plants and animals of the Pleistocene are, in many respects, similar to those living today, but important differences exist. Moreover, the spatial distribution of various Pleistocene fauna and flora types differed markedly from what it is at present. Changes in climate and environment caused large-scale migrations of both plants and animals, evolutionary adaptations, and in some cases extinction. Study of the biota provides not only data on the past paleoenvironments but also insights into the response of plants and animals to well-documented environmental change. Of particular importance is the evolution of the genus *Homo* during the Pleistocene and the extinction of large mammals at the end of the epoch.

Evolutionary changes. Evolutionary changes during the Pleistocene generally were minor because of the short interval of time involved. They were greatest among the mammals. In fact, the epoch has been subdivided into mammalian ages on the basis of the appearance of certain immigrant or endemic forms (Table 27).

Table 27: Pleistocene Mammalian Ages in North America and Europe

North America	Europe
Rancholabrean	Steinheimian
Irvingtonian	Biharian
Blancan*	Villanyian*
	Villafranchian*

*These mammalian ages are Pliocene, but the associated forms are relatively similar to those of Pleistocene faunas.

Mammalian evolution included the development of large forms, many of which became adapted to Arctic conditions. Among these were the woolly mammoth, woolly rhinoceros, musk ox, moose, reindeer, and others that inhabited the cold periglacial areas. Large mammals that inhabited the more temperate zones included the elephant, mastodon, bison, hippopotamus, wild hog, deer, giant beaver, horse, and ground sloth. The evolution of these as well as of much smaller forms was affected in part by three factors: (1) a generally cooler, more arid climate subject to periodic fluctuations, (2) new migration routes resulting largely from the emergence of intercontinental connections during times of lower sea level, and (3) a changing geography due to the uplift of plateaus and mountain building.

Appearance of hominids

The most significant biological development was the appearance and evolution of the genus *Homo*. The oldest species, *H. habilis,* probably evolved from an australopithecine ancestor in the late Pliocene. The species was present in Africa by 2 million years ago and is known from sites as young as 1.5 million years old. Another extinct species, *H. erectus,* evolved in Africa, possibly from *H. habilis,* and is known from sites about 1.6 million years old. *H. erectus* spread to other parts of the Old World during the early Pleistocene and is known from northern China and Java by roughly 1 million years ago. Representatives of this group are known from many sites, and these beings constituted the dominant human species for more than a million years. The species *H. sapiens,* to which all modern humans belong, evolved in the later part of the middle Pleistocene, and early forms of the species are known from about 400,000 years ago. More modern forms of *H. sapiens,* the Neanderthals, appeared approximately 100,000 years ago during the last interglaciation and are known from many sites in Europe and western Asia. They disappeared about 35,000 to 30,000 years ago, and by then populations with fully modern skeletons had evolved and were widespread in the Old World. Exactly when modern *H. sapiens* entered the New World remains controversial. It appears that fully evolved humans had migrated as far as Alaska from Siberia via the Bering land bridge by 30,000 years ago, and large numbers presumably moved south down the Canadian plains corridor between the Cordilleran and Laurentide ice sheets when it opened near the end of the last glaciation some 12,000 years ago. Conflicting and not fully accepted evidence at a few sites in the United States and in southern South America, however, suggests occupation of the continental interior prior to 30,000 years ago. If such findings are valid, the group of earlier immigrants may have arrived by small ocean-going craft from the Pacific Islands.

Migration of plants and animals. Changing environments in response to climatic variation caused drastic disruptions of faunas and floras both on land and in the oceans. These disruptions were greatest near the former ice sheets that extended far to the south and caused the southward displacement of climatic and vegetation zones. In the temperate zones of central Europe and the United States where deciduous forests exist today, vegetation was open and most closely resembled the northern tundra, with grasses, herbs, and few trees during glacial intervals. Farther south, a broad region of boreal forests with varying proportions of spruce and pine or a combination of both extended almost to the Mediterranean in Europe and northern Louisiana in North America. The vegetation succession has been documented by studies of fossil pollen, which accumulated year by year with other sediments in lakes and bogs beyond the ice margin (see Figure 46). Although such floral migrations appear simple in concept, interpretation of the vegetation record is quite complicated because a number of the glacial pollen assemblages have no modern analogues—*i.e.,* they contain mixtures of forms from different present-day climatic environments. Similar relationships also occur with vertebrate faunas: more temperate forms commonly occur together with more Arctic forms. Such "disharmonious" faunas suggest that glacial climatic and environmental conditions in some cases were totally unlike those of any modern environment. One explanation is that climatic conditions may have been more equable during glacial times and may have lacked the seasonal extremes of modern climates in such areas. Although overall temperatures were significantly lower, summers probably were much cooler because of the influence of the ice sheet, and winters, except very near the ice margin, lacked severe cold spells, as the ice sheet formed a barrier to Arctic air masses that today bring freezing conditions far to the south. Thus, plants and animals whose geographic ranges would ordinarily be controlled by either

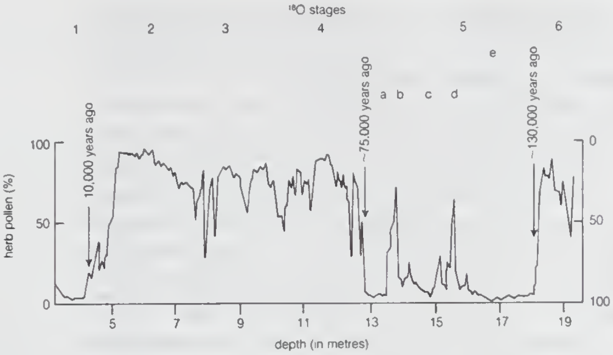


Figure 46: Relative proportions of herb and tree pollen from a core of a peat bog at Grande Pile, northwestern France. Correlative marine oxygen isotope stages are indicated. Vegetation varied from deciduous forest to tundra with abrupt changes in vegetation.

From G. Woillard (1979) in W S Broecker and G H Denton, *Geochimica et Cosmochimica Acta*, vol 33, © 1989 Pergamon Press

extreme seasonal warm or cold conditions were able to coexist during glacial times, and considerable community reorganization took place in response to climatic change during and following a glaciation.

Similar responses to changing environments are well known from life in the oceans. Marine organisms closely reflect the temperature, depth, and salinity of the water in which they live, and studies of the fossil succession from deep-sea cores have allowed detailed reconstructions of oceanic conditions for the late Pleistocene. Planktonic foraminifers are most useful for determining sea-surface conditions, and changes in the distribution of polar, subpolar, subtropical, and tropical faunas have been used to map changing oceanic conditions. Changes in the North Atlantic Ocean were most dramatic because of the direct influence of the ice sheets to the west, north, and east. During episodes of glaciation, polar faunas extended south to about 45° N latitude, whereas during interglaciations these faunas occurred mostly north of 70° and subtropical faunas extended far to the north under the influence of the Gulf Stream.

Displacement of marine forms

Megafaunal extinctions. The end of the Pleistocene was marked by the extinction of many genera of large mammals, including mammoths, mastodons, ground sloths, and giant beavers. The extinction event is most distinct in North America, where 32 genera of large mammals vanished during an interval of about 2,000 years, centred on 11,000 BP. On other continents, fewer genera disappeared, and the extinctions were spread over a somewhat longer time span. Nonetheless, they still appear to be more common near the end of the Pleistocene than at any other time during the epoch. Except on islands, small mammals, along with reptiles and amphibians, generally were not affected by the extinction process. The cause of the extinctions has been vigorously debated, with two main hypotheses being advanced: (1) the extinctions were the result of overpredation by human hunters; and (2) they were the result of abrupt climatic and vegetation changes during the last glacial–interglacial transition.

The first theory, the so-called overkill hypothesis, receives support from the coincidence in the timing of the mass extinction and the appearance of large numbers of human hunters, as evidenced by the Clovis complex, an ancient culture centred in North America. Clovis archaeological sites (concentrated in Arizona, New Mexico, and West Texas), with their distinctive projectile points, date between 10,000 and 12,000 years ago. Proponents of the hypothesis point out that these new immigrants from Eurasia were skilled hunters, that the North American fauna would not have been wary of this new group of predators, and that, once the number of large herbivores declined, large carnivores also would have been affected as their prey became extinct. In addition to direct slaughter, human disruption of the environment most likely contributed to the extinctions, particularly on other continents.

Abrupt climatic change also occurred at the time of the megafaunal extinctions, and so timing alone does not clearly differentiate one hypothesis from the other. The climatic-change hypothesis takes a number of forms but essentially focuses on the reorganization of vegetation, on the availability of food (including nutrient value), and on the general environmental disruption and stress that resulted as climates became more seasonal. It appears likely that the causes of extinction varied in different geographic areas under different conditions and that both climatic change and human activities played roles but of varying importance in different situations.

Cause of the climatic changes and glaciations. Pleistocene climates and the cause of the climatic cycles that resulted in the development of large-scale continental ice sheets have been a topic of study and debate for more than 100 years. Many theories have been proposed to account for Quaternary glaciations, but most are deficient in view of current scientific knowledge about Pleistocene climates. One early theory, the theory of astronomical cycles, seems to explain much of the climatic record and is considered by most to best account for the fundamental cause or driving force of the climatic cycles.

Astronomical theory

The astronomical theory is based on the geometry of the Earth's orbit around the Sun, which affects how solar radiation is distributed over the surface of the planet. The latter is determined by three orbital parameters that have cyclic frequencies: (1) the eccentricity of the Earth's orbit (*i.e.*, its departure from a circular orbit), with a frequency of about 100,000 years, (2) the obliquity, or tilt, of the Earth's axis away from a vertical drawn to the plane of the planet's orbit, with a frequency of 41,000 years, and (3) the precession, or wobble, of the Earth's axis, with frequencies of 19,000 and 23,000 years. Collectively these parameters determine the amount of radiation received at any latitude during any season; radiation curves have been calculated from them for different latitudes for the past 600,000 years. These curves vary systematically from the poles to the equator, with those in the higher latitudes being dominated by the 41,000-year tilt cycle and those in lower latitudes by the 19,000- and 23,000-year precession cycles. The astronomical theory places emphasis on summer insolation in the high-latitude areas of the Northern Hemisphere (about 55° N latitude). Glaciations are hypothesized to begin during times of low summer insolation when conditions should be most optimal for winter snow to last through the summer season.

Dating of the marine terraces in Barbados and New Guinea and, more importantly, determining the chronology of glaciations as inferred from the marine oxygen isotope record were milestones in testing the astronomical theory. Early spectral analysis of the oxygen isotope record of cores from the deep ocean showed frequencies of climatic variation at essentially the same frequencies as the orbital cycles—that is to say, at 100,000 years, 43,000 years, 24,000 years, and 19,000 years. These results (reported in 1976), along with those of more recent analyses, provide firm evidence of a tie between orbital cycles and the Earth's recent climatic record. The variations in the Earth's orbit are generally considered the "pacemaker" of the ice ages.

Although the planetary orbital cycles are the likely cause of the Pleistocene climatic cycles, the mechanisms and connections to the global climate are not fully understood, and important questions remain unanswered. The relatively small seasonal and latitudinal radiation variations alone cannot account for the magnitude of climatic change as experienced by the Earth during the Pleistocene. Clearly, feedback mechanisms must operate to amplify the insolation changes caused by the orbital parameters. One of these is albedo, the reflectivity of the Earth's surface. Increased snow cover in high-latitude areas would cause increased cooling. Another feedback mechanism is the decreased carbon dioxide content of the atmosphere during times of glaciation, as recorded in the bubbles of long ice cores. Variations in atmospheric carbon dioxide are essentially synchronous with global climatic change and thus in all likelihood played a significant role through the so-called greenhouse effect. (The latter phenomenon refers to the trapping of heat—that is to say, infrared radiation—in the lower levels of the atmosphere by carbon dioxide, water vapour, and certain other gases.) Another atmospheric effect is the increased amount of dust during glacial times, as borne out by ice core and loess records. All of these changes operate in the same direction, causing increased cooling during glacial times and warming during interglacial times.

Other problems remain with respect to the astronomical theory. One is the dominance of the 100,000-year cycle in the Pleistocene climatic record, whereas the eccentricity cycle is the weakest among the orbital parameters. Another is the cause of the asymmetrical pattern of the climatic record. Ice ages appear to start slowly and take a long time to build up to maximum glaciation, only to terminate abruptly and go from maximum glacial to full interglacial conditions in less than 10,000 years (see Figure 41). A third problem is the synchronous nature of the climatic record between the Northern and Southern hemispheres, which one would not expect from the orbital parameters because they operate in different directions in the two hemispheres.

Different approaches have been taken to explain these questions. Most of these suggest that the Northern Hemisphere with its enormous continental ice sheets was the controlling area and that the ice sheets themselves with their complex dynamics may explain the 100,000-year climatic cycle. Others propose that major reorganizations of the ocean–atmosphere system must be called upon to explain the climatic record. These reorganizations are concerned with the transport of salt through the oceans and water vapour through the atmosphere and revolve around the existence and strength of deep oceanic currents in the Atlantic Ocean.

Ongoing interdisciplinary research on Pleistocene paleoclimatology is focused on understanding the complex dynamics and interactions among the atmosphere, oceans, and ice sheets. Such research is expected to provide further insight into the cause of the climatic cycles, which is essential as scientists attempt to predict future climates in view of recent human-induced modifications of the climatic system. (W.H.J.)

HOLOCENE EPOCH

General considerations. The Holocene, also referred to as the Recent, is the latest interval of geologic time, covering approximately the last 10,000 years of Earth history (see Table 4). The sediments of the Holocene, both continental and marine, cover the largest area of the globe of any epoch in the geologic record, but the Holocene is unique because it is coincident with the late and post-Stone Age history of mankind. The influence of humans is of world extent and is so profound that it seems appropriate to have a special geologic name for this time. In 1833 Charles Lyell proposed the designation Recent for the period that has elapsed since "the earth has been tenanted by man." It is now known that humans have been in existence a great deal longer. The term Holocene was proposed in 1867 and was formally submitted to the International Geological Congress at Bologna, Italy, in 1885. It was officially endorsed by the U.S. Commission on Stratigraphic Nomenclature in 1969.

The Holocene is the latest division of the Quaternary and represents the most recent interglacial interval of the period. The preceding and substantially longer sequence of alternating glacial and interglacial ages is the Pleistocene Epoch (see above). Because there is nothing to suggest that the Pleistocene has actually ended, certain authorities prefer to extend the Pleistocene up to the present time; this approach tends to ignore humans and their impact, however. The Holocene forms the chronological framework for human history. Archaeologists use it as the time standard against which they trace the development of early civilizations.

Radiocarbon dating

Stratigraphy. *Chronology and correlation.* The Holocene is unique among geologic epochs because varied means of correlating deposits and establishing chronologies are available. One of the most important means is carbon-14 dating (see above *Carbon-14 dating and other cosmogenic methods*). Because the age determined by the carbon-14 method may be appreciably different from the true age in certain cases, it is customary to refer to such dates in "radiocarbon years." These dates, obtained from a variety of deposits, form an important framework for Holocene stratigraphy and chronology.

The limitations of accuracy of radiocarbon age determinations are expressed as ± a few tens or hundreds of years. In addition to this calculated error, there also is a question of error due to contamination of the material measured. For instance, an ancient peat may contain some younger roots and thus give a falsely "young" age unless it is carefully collected and treated to remove contaminants. Marine shells consist of calcium carbonate ($CaCO_3$), and in certain coastal regions there is upwelling of deep oceanic water that can be 500 to more than 1,000 years old. An "age" from living shells in such an area can suggest that they are already hundreds of years old.

Table 28 shows the comparative dates of radiocarbon years and those obtained by other means. Two sets of radiocarbon years are given because the half-life of carbon-14 was reassigned a value of 5,730 years by agreement of scientists. Many dates available in the literature, however, are based on the originally established half-life of 5,570 years.

Table 28: Comparative Dating Systems for the Holocene Epoch*
(and latest Pleistocene)

uncorrected radiocarbon BP dates (radiocarbon years)		U.S. tree-ring dates in absolute years BP (before present, AD 1950, in sidereal years, adjusted to radiocarbon according to Damon)‡	AD/BC dates (sidereal years)	conventional BP varve years (estimated to be 350 ± 200 years too young)§
(T_2=5570)†	(T_2=5730)†			
1000	1000	900	AD 1050	...
2000	2050	1950	AD 0	(1600)
3000	3100	3250	1300 BC	(2900)
4000	4150	4650	2700	(4300)
5000	5180	5920	3970	(5630)
6000	6200	6900	4950	(6550)
7000	7220	(7450)	5550	7100
8000	8240	(8350)	6400	8000
9000	9270	(9200)	7250	8850
10,000	10,300	(10,550)	7600	10,200
11,000	11,330	(11,550)	9600	11,200
12,000	12,350	(12,550)	10,600	12,200

*Wood from the tomb of the pharaoh Djoser at Saqqārah, Egypt, is dated historically at 4650 ± 75 BP in sidereal years, but according to multiple analyses by many laboratories it is about 4100 BP in radiocarbon years, or 550 years too young. The anomaly is most probably explained in terms of solar radiation, residence time of CO_2, and paleomagnetics. Tree-ring dating combined with ^{14}C measurements has confirmed this trend and provides a general curve for correcting Holocene ^{14}C dates. Almost all radiocarbon dates given are uncorrected. U.S. tree-ring chronology most closely resembles the sidereal dates; the Scandinavian varve chronology also is close to the astronomical chronology, subject to a 350 ± 200 year correction. †T_2 signifies half-life. ‡Dates older than 7450 BP are based on varve years, corrected by 350 ± 200 years. §Dates for varve years less than 6550 BP are extrapolated.

Varve chronology

In certain areas a varve chronology can be established. This involves counting and measuring thicknesses in annual paired layers of lake sediments deposited in lakes that undergo an annual freeze-up. Because each year's sediment accumulation varies in thickness according to the climatic conditions of the melt season, any long sequence of varve measurements provides a distinctive "signature" and can be correlated for moderate distances from lake basin to lake basin. The pioneer in this work was the Swedish investigator Baron Gerard De Geer, who developed a long chronology on which that shown in Table 28 is partly based.

In some relatively recent continental deposits, obsidian (a black glassy rock of volcanic origin) can be used for dating. Obsidian weathers slowly at a uniform rate, and the thickness of the weathered layer is measured microscopically and gauged against known standards to give a date in years. This has been particularly useful where arrowheads of obsidian are included in deposits.

As noted elsewhere in this article, paleomagnetism is another phenomenon used in chronology. The Earth's magnetic field undergoes a secular shift that is fairly well known for the last 2,000 years. The magnetized material to be studied can be natural, such as a lava flow; or it may be man-made, as, for example, an ancient brick kiln or smeltery that has cooled and thus fixed the magnetic orientation of the bricks to correspond to the geomagnetic field of that time.

Another form of dating is tephrochronology, so called because it employs the tephra (ash layers) generated by volcanic eruptions. The wind may blow the ash 1,500–3,000 kilometres, and, because the minerals or volcanic glass from any one eruptive cycle tend to be distinctive from those of any other cycle, even from the same volcano, these can be dated from the associated lavas by stratigraphic methods (with or without absolute dating). The ash layer then can be traced as a "time horizon" wherever it has been preserved. When the Mount Mazama volcano in Oregon exploded at about 6600 BP (radiocarbon-dated by burned wood), 70 cubic kilometres of debris were thrown into the air, forming the basin now occupied by Crater Lake. The tephra were distributed over 10 states, thereby providing a chronological marker horizon. A comparable eruption of Thera on Santorin in the Aegean Sea about 3,400 years ago left tephra in the deep-sea sediments and on adjacent land areas. Periodic eruptions of Mount Hekla in Iceland have been of use in Scandinavia, which lies downwind.

Tree-ring measurement and analysis

Finally, the measurement and analysis of tree rings (or dendrochronology) must be mentioned. The age of a tree that has grown in any region with a seasonal contrast in climate can be established by counting its growth rings. Work in this field by the University of Arizona's Laboratory of Tree-Ring Research, by selection of both living trees and deadwood, has carried the year-by-year chronology back more than 7,500 years. Certain pitfalls have been discovered in tree-ring analysis, however. Sometimes, as in a very severe season, a growth ring may not form. In certain latitudes the tree's ring growth correlates with moisture, but in others it may be correlated with temperature. From the climatic viewpoint these two parameters are often inversely related in different regions. Nevertheless, in experienced hands, just as with varve counting from adjacent lakes, ring measurements from trees with overlapping ages can extend chronologies back for many thousands of years. The bristlecone pine of the White Mountains in California has proved to be singularly long-lived and suitable for this chronology; some individuals still living are more than 4,000 years old, certainly the oldest living organisms. Wood from old buildings and even old paving blocks in western Europe and in Russia have contributed to the chronology. This technique not only offers an additional means of dating but also contains a built-in documentation of climatic characteristics. In certain favourable situations, particularly in the drier, low latitudes, tree-ring records sometimes document 11- and 22-year sunspot cycles.

The Pleistocene–Holocene boundary. Arguments can be presented for the selection of the lower boundary of the Holocene at several different times in the past. Some Russian investigators have proposed a boundary at the beginning of the Allerød, a warm interstadial age that began about 12,000 BP. Others, in Alaska, proposed a Holocene section beginning at 6000 BP. Marine geologists have recognized a worldwide change in the character of deep-sea sedimentation about 10,000–11,000 BP. In warm tropical waters, the clays show a sharp change at this time from chlorite-rich particles often associated with fresh feldspar grains (cold, dry climate indicators) to kaolinite and gibbsite (warm, wet climate indicators).

Pleistocene–Holocene boundary dates

Some of the best-preserved traces of the boundary are found in southern Scandinavia, where the transition from the latest glacial stage of the Pleistocene to the Holocene was accompanied by a marine transgression. These beds, south of Göteborg, have been uplifted and are exposed at the surface. The boundary is dated around 10,300 ± 200 years BP (in radiocarbon years). This boundary marks the very beginning of warmer climates that occurred after the latest minor glacial advance in Scandinavia. This advance built the last Salpausselkä moraine, which corresponds in part to the Valders substage in North America. The subsequent warming trend was marked by the Finiglacial retreat in northern Scandinavia, the Ostendian (early Flandrian) marine transgression in northwestern Europe.

Nature of the Holocene record. The very youthfulness of the Holocene stratigraphic sequence makes subdivision difficult. The relative slowness of the Earth's crustal movements means that most areas which contain a complete marine stratigraphic sequence are still submerged. Fortunately, in areas that were depressed by the load of glacial ice there has been progressive postglacial uplift (crustal rebound) that has led to the exposure of the nearshore deposits.

Character and fossil content of globigerina ooze

Deep oceanic deposits. The marine realm, apart from covering about 70 percent of the Earth's surface, offers far better opportunities than coastal environments for undisturbed preservation of sediments. In deep-sea cores, the boundary usually can be seen at a depth of about 10–30 centimetres, where the Holocene sediments pass downward into material belonging to the late glacial stage of the Pleistocene. The boundary often is marked by a slight change in colour. For example, globigerina ooze, common in the ocean at intermediate depths, is frequently slightly pinkish when it is of Holocene age because of a trace of iron oxides that are characteristic of tropical soils. At greater depth in the section, the globigerina ooze may be grayish because of greater quantities of clay, chlorite, and feldspar that have been introduced from the erosion of semiarid hinterlands during glacial time.

During each of the glacial epochs the cooling of the ocean waters led to reduced evaporation and thus fewer clouds, then to lower rainfall, then to reduction of vegetation, and so eventually to the production of relatively more clastic sediments (owing to reduced chemical weathering). Furthermore, the worldwide eustatic (glacially related) lowering of sea level caused an acceleration of erosion along the lower courses of all rivers and on exposed continental shelves, so that clastic sedimentation rates in the oceans were higher during glacial stages than during the Holocene. Turbidity currents, generated on a large scale during the low sea-level periods, became much less frequent following the rise of sea level in the Holocene.

Studies of the fossils in the globigerina oozes show that at a depth in the cores that has been radiocarbon-dated at about 10,000–11,000 BP the relative number of warm-water planktonic foraminiferans increases markedly. In addition, certain foraminiferal species tend to change their coiling direction from a left-handed spiral to a right-handed spiral at this time. This is attributed to the change from cool water to warm water, an extraordinary (and still not understood) physiological reaction to environmental stress (Figure 47). Many of the foraminiferans, however, responded to the warming water of the Holocene by migrating poleward by distances of as much as 1,000 to 3,000 kilometres in order to remain within their optimal temperature habitats.

Nannoplankton and oceanic circulation

In addition to foraminiferans in the globigerina oozes, there are nannoplankton, minute fauna and flora consisting mainly of coccolithophores. Research on the present coccolith distribution shows that there is maximum pro-

Adapted from G Wollin, D Ericson and M Ewing 'Late Pleistocene Climates Recorded in Atlantic and Pacific Deep-Sea Sediments,' in *Late Cenozoic Glacial Ages* (1971) Yale University Press, New Haven and London

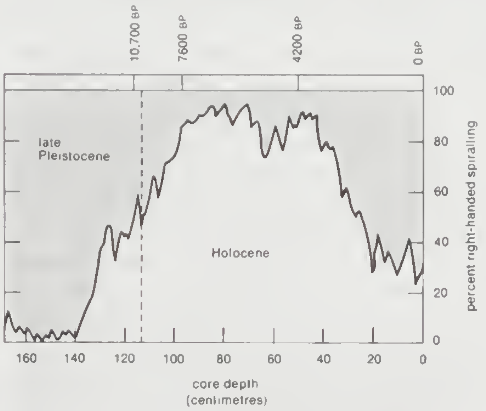


Figure 47: Climatic curve based on coiling direction of *Globoratalia truncatulinoides* obtained from deep-sea core at latitude 24°18′ N, longitude 75 55′ W (see text).

ductivity in zones of oceanic upwelling, notably at the subpolar convergence and the equatorial divergence. During the latest glacial stage the subpolar zone was displaced toward the equator, but with the subsequent warming of waters it shifted back to the borders of the polar regions.

The distribution of the carbonate plankton bears on the problem of rates of oceanic circulation. Is the Holocene rate higher or lower than during the last glacial stage? It has been argued that, because of the higher mean temperature gradient in the lower atmosphere from equator to poles during the last glacial period, there would have been higher wind velocities and, because of the atmosphere–ocean coupling, higher oceanic current velocities. There were, however, two retarding factors for glacial-age currents. First, the eustatic withdrawal of oceanic waters from the continental shelves reduced the effective area of the oceans by 8 percent. Second, the greater extent of floating sea ice would have further reduced the available air–ocean coupling surface, especially in the critical zone of the westerly circulation. According to climatic studies by the British meteorologist Hubert H. Lamb, the presence of large continental ice sheets in North America and Eurasia would have introduced a strong blocking action to the normal zonal circulation of the atmosphere, which then would be replaced by more meridional circulation. This in turn would have been appreciably less effective in driving major oceanic current gyres.

Continental shelf and coastal regions. It was recognized as early as 1842 that a logical consequence of a glacial age would be a large-scale withdrawal of ocean water. Consequently, deglaciation would produce a postglacial "glacioeustatic" transgression of the seas across the continental shelf. The trace of this Holocene rise of sea level was first discerned along the New England coast and along the coast of Belgium, where it was named the Flandrian Transgression by Georges Dubois in 1924.

Whereas the deep-sea Holocene sediments usually follow without interruption upon those of the Upper Pleistocene, on the continental shelf there is almost invariably a break in the sequence upon the continental formations there. As sea level rose, it paused or fluctuated at various stages, leaving erosional terraces, beach deposits, and other indicators of the stillstand. Brief regressions in particular permitted the growth of peat deposits that are of significance in the Holocene record because they can be dated by radiocarbon analysis. Dredging in certain places on the shelf, such as off eastern North America, also is useful because terrestrial fossils from the latest glacial period or early Holocene have been found; these range from mammoth and mastodon bones and tusks to human artifacts.

Sedimentary deposits and coral reefs

On about 70 percent of the world's continental shelves today the amount of sedimentary accumulation since the beginning of the Holocene is minimal, so that dredging or coring operations often disclose hard rock, with older formations at or very close to the surface. In other places, especially near the former continental ice fronts, the shelf is covered by periglacial fluvial sands (meltwater deposits), which, because of their unconsolidated nature, became extensively reworked into beaches and bars during the Holocene Transgression.

In warm coral seas the major pauses in the Holocene eustatic rise were long enough for fringing reefs to become established; and, when the rise resumed, the reefs grew upward, either in ribbonlike barriers or from former headlands as patch reefs or shelf atolls. Since coral generally does not colonize a sediment-covered shelf floor at depths of more than about 10 metres, those reefs now rising from greater depths must have been emplaced in the early Holocene or grown on foundations of ancient reefs.

The great ice-covered areas of the Quaternary Period included Antarctica, North America, Greenland, and Eurasia. Of these, Antarctica and Greenland have relatively high latitude situations and do not easily become deglaciated. Some melting occurs, but there is a very great melt-retardation factor in high-latitude ice sheets (high albedo or reflectivity, short melt season, and so forth). In the case of mid-latitude ice sheets, however, once melting starts, the ice disappears at a tremendous rate. The melt rate reached a maximum about 8000 BP, liberating 18 trillion (18×10^{12}) metric tons of meltwater annually. This corresponds to a rise in sea level of five centimetres per year. Hand in hand with melting, the sea level responded so that, as the ice began to retreat from its former terminal moraines, the sea began to invade the former coastlands.

As the sea level rose, the Earth's crust responded buoyantly to the removal of the load of ice, and at critical times the rate of rise of the water level was outstripped by the rate of rise of the land. In these places the highest ancient shoreline that is now preserved is known as the marine limit. The nearer the former centre of the ice sheet, the higher the marine limit. In northern Scandinavia, Ontario and northwestern Quebec, around Hudson Bay, and in Baffin Island, it reaches more than a 300-metre elevation. In central Maine and Spitsbergen it may exceed 100 metres, whereas in coastal Scotland and Northern Ireland it is rarely above 10–15 metres.

In addition to the marine-limit strandlines, there are row upon row of lower beach levels stretched out across Scandinavia, around Hudson Bay, and on other Arctic coasts. These strandlines are dated and distinctive and do not grade into each other. Each represents a specific period of time when the rising crust and rising sea level remained in place long enough to permit the formation of beaches, spits, and bars and sometimes the erosion of headlands ("fossil cliffs").

A complicating factor near the periphery of former ice sheets is the so-called marginal bulge. Reginald A. Daly, an American geologist, postulated that, if the ice load pressed down the middle of the glaciated area, then the Earth's crust in the marginal area tended to rise up slightly, producing a marginal bulge. With deglaciation the marginal bulge should slowly collapse. A fulcrum should develop between postglacial uplift and peripheral subsidence. In North America that fulcrum seems to run across Illinois to central New Jersey and then to swing northeastward, paralleling the coast and turning seaward north of Boston. In the Scandinavian region the fulcrum crosses central Denmark to swing around the Baltic Sea and then trends northeastward across the Gulf of Finland north of St. Petersburg, so that the southeastern Baltic and northwestern Germany are subsiding. The Netherlands area is subsiding also, but here the pattern is complicated by the long-term negative tectonic trend of the North Sea Basin and the Rhine delta.

It seems likely that this fulcrum shifted inward toward the former glacial centre during the early part of the Holocene. Passing inland, the lines of equal uplift (isobases) are positive, whereas seaward they are negative. The coastal area of southern New England is still slowly subsiding at the present time (1–3 millimetres per year).

The great deltas of the world, those of the Mississippi, Rhine, Rhône, Danube, Nile, Amazon, Niger, Tigris-Euphrates, Ganges, and Indus, all coincide with regions of tectonic subsidence. Because water-saturated sediment has a tendency to compact under further sediment loading, there is an additional built-in mechanism that adds to the subsidence in such areas.

Deltaic environment and sedimentary facies

In this deltaic setting Holocene sequences are found that are quite different from those in the postglacial uplifted regions. Whereas the Holocene beaches in the uplift areas extend horizontally across the country in concentric belts, the Holocene sequence in the deltaic regions is, for the most part, vertical in nature and can be studied only from well data.

In both the Mississippi and Rhine deltas, sediments that represent the earliest marine Holocene are missing. The sediments must lie seaward on the shelf margin, and the oldest marine layers are found to rest directly upon the late Pleistocene river silts and gravels. In a delta settling at around 0.5 to 3 millimetres per year, the rising sea of the Flandrian Transgression extended quickly across the river deposits to the inner margin (where there is a fulcrum comparable to that of the glaciated regions), marking the boundary between areas of downwarp and those of relative stability or gentle upwarp. The marine beds alternate with continental deposits that represent river or swamp environments. Six major fluctuations are recognizable in both the Mississippi and Rhine deltas. By radiocarbon

Table 29: Peat Building, Sea-Level Fluctuation, and the Holocene Record of the Mississippi Delta

Mississippi stage	years BP	international correlation (regression)	probable eustatic range
Balize Delta	300	Late Medieval	0 to −0.5 m
Plaquemines Delta	700	Paria	+0.5 to −1.5 m
Lafourche Delta	1500	pre-Dunkerquian III	0 to −1 m
Peat #1 (St. Bernard or La Loutre Delta)	1700–2100	Florida (Roman)	+1.5 to −2 m
Teche Delta	2800–3300	Pelham Bay	+2 to −2 m
Peat #2 (Cocodrie Delta)	4000–4300	Bahama	? +3 to −3 m
Peat #3 (Marinquin, Sale, or Cypremort Delta)	4700–5000	early Subboreal	? +3 to −3 m
Peat #4	6000–6500	Rhine Delta	−4 to −9 m
Peat #5	7000–7500	late Boreal	−12 to −22 m

Source: Partly from Kolb and Van Lopik, 1966, and Fairbridge, 1968.

dating the transgressive and regressive phases (Table 29) have been shown to be correlative in time.

Beaches, dunes, and peat deposits

On a subsiding coast there tends to be an alternation in importance between two types of associated sedimentary facies. During a regression of the sea the river distributaries are rejuvenated and there is an increase in the supply of sand and silt; beaches are widened and beach ridge dunes or cheniers may be formed. During a transgressive stage the saltwater wedge at river mouths causes a back-up, and the estuary becomes much more sluggish (thalassostatic).

In The Netherlands the basal Holocene is buried in the fluvial deposits of the lower Rhine. The postglacial eustatic rise had to traverse the North Sea Plain and advance up the English Channel several hundred kilometres before it reached the Netherlands area. At about 9000–8500 BP (Ancylus stage in the Baltic), the coastal beaches still lay seaward from the present shore. Subsequently, they became stabilized by a brief eustatic regression, while the high water table permitted the growth of the Lower Peat. This is contemporaneous with the late Boreal Peat that is widespread in northern Europe, as well as Peat #5 of the Mississippi delta (Table 29).

A further eustatic rise (of about 10–12 metres) ensued around 7750 BP, corresponding to a warming of the climate marked by the growth of oak forests in western Europe (the BAT, or "Boreal–Atlantic Transition"). In The Netherlands the barrier beaches re-formed close to the present coastline, and widespread tidal flats developed to the interior. These are known as the Calais Beds (or Calaisian) from the definition in Flanders by Dubois. In the protected inner margins, the peat continued to accumulate during and after the "Atlantic" time.

From evidence outside the areas of subsidence, it seems likely that the worldwide eustatic sea level rise reached its maximum sometime between 5500 and 2500 BP (many workers consider the date to be about 2000 BP). In The Netherlands, in spite of subsidence, the western coastline became more or less stabilized around 4000 BP with the beginning of the formation of the Older Dunes alternating with interdune soils. At the same time, in the tide flat areas the Calaisian was followed by the Dunkirk stage, or Dunkerquian.

The Younger Dune sequence of The Netherlands began with a dry climatic phase in the 12th century AD. With several fluctuations of cold continental climates, dune building continued until the 16th century. Only brief positive oscillations of sea level occurred until the 17th century, when the "modern" warming and eustatic rise started, accompanied also by dune stabilization.

Broadly comparable patterns occur in other areas, from France and Britain to Texas, Oregon, and Brazil. There is normally a threefold or fourfold subdivision in all the Holocene coastal dune belts, each extensively vegetated and consolidated before the successively younger dune belt was added. In a number of cases there is evidence from buried beach deposits that the foundations of the inner dunes are older strandlines that were established when the sea was somewhat higher than today. An important regressive phase seems to have initiated each new dune belt.

Other coastal regions. Besides regions of glacio-isostatic crustal adjustment, both positive and negative, and the deltaic or geosynclinally subsiding areas, there are many tens of thousands of kilometres of coastlines that are relatively stable and a smaller fraction that are tectonically active.

Holocene terraces and the eustatic curve

Most striking scenically are the coasts with Holocene terraces undergoing tectonic uplift. Terraces of this sort, backed in successive steps by Pleistocene terraces, are well developed in South America, the East Indies, New Guinea, and Japan. By careful surveys every few years the Japanese geodesists have been able to establish mean rates of crustal uplift (or subsidence) for many parts of the country and have been able to construct a residual eustatic curve that is comparable with those obtained elsewhere.

Besides uplifted coasts outside of glaciated areas there are also certain highly indented coasts that show clear evidence of Holocene "drowning." These coasts typically are characterized by the rias, or drowned estuaries, sculptured by fluvial action, but many of the valleys were cut 10 to 20 million years ago, and the Holocene history has been purely one of eustatic rise.

On the basis of the known climatic history of the Holocene, from the strandline record of Scandinavia and from the sedimentologic evolution of the Mississippi and Rhine deltas, an approximate chronology of Holocene eustasy can be worked out. The amplitudes of the fluctuations and the finite curve are less easily established. A first approximation of the oscillations was published in 1959 and in a more detailed way in 1961 (the so-called Fairbridge curve). Smoothed versions have been offered by several other workers (see Figure 48).

Holocene environment and biota. In formerly glaciated regions, the Holocene has been a time for the reinstitution of ordinary processes of subaerial erosion and progressive reoccupation by a flora and fauna. The latter expanded rapidly into what was an ecological vacuum, although with a very restricted range of organisms, because the climates were initially cold and the soil was still immature.

Floral change. The most important biological means of establishing Holocene climate involves palynology, the study of pollen, spores, and other microscopic organic particles. Pollen from trees, shrubs, or grasses is generated annually in large quantities and often is well preserved in fine-grained lake, swamp, or marine sediments. Statistical correlations of modern and fossil assemblages provide a

Adapted from N. Morner, *Sveriges Geologiska Undersokning*, Series C, no. 640 (1969)

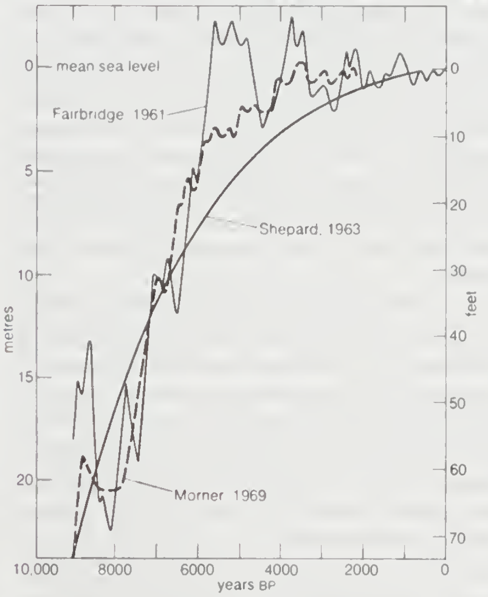


Figure 48: *Holocene sea level curves.*
The smooth curve of Shepard represents an average of many radiocarbon-dated shoreline indicators from both stable and subsiding areas; the oscillating curve of Mörner is based upon the isostatically emerged coast of western Sweden, corrected for uplift; and the oscillating curve of Fairbridge is based upon the geomorphology and radiocarbon dates of coastal features of the world, adjusted for crustal movements.

Table 30: Blytt–Sernander Framework

Blytt and Sernander divisions	archaeological periods in Scandinavia		absolute chronology|| years, sidereal	radiocarbon years, uncorrected BP (1950)
Sub-Atlantic	Historical time		AD 1950	0
	Viking time		AD 1000	1,050
	Iron Age		0	2,000
			600 BC	2,600
Subboreal	Bronze Age			
			1900 BC	3,600
	Neolithic Stone Age			
			4100 BC	5,100
Atlantic	Mesolithic Stone Age	Ertebølle Industry	6150 BC	7,750
Boreal		Maglemosian Industry	7100 BC	9,500
Preboreal		Klosterlund Industry	7700 BC	10,100

basis for estimating the approximate makeup of the local or regional vegetation through time. Even a crude subdivision into arboreal pollen (AP) and nonarboreal pollen (NAP) reflects the former types of climate. The tundra vegetation of the last glacial epoch, for example, provides predominantly NAP, and the transition to forest vegetation shows the climatic amelioration that heralded the beginning of the Holocene.

The first standard palynological stratigraphy was developed in Scandinavia by Axel Blytt, Johan Rutger Sernander, and E.J. Lennart von Post, in combination with a theory of Holocene climate changes. The so-called Blytt–Sernander system was soon tied to the archaeology and to the varve chronology of Gerard De Geer. It has been closely checked by radiocarbon dating, establishing a very useful standard. Every region has its own standard pollen stratigraphy, but these are now correlated approximately with the Blytt–Sernander framework (Table 30). To some extent this is even true for remote areas such as Patagonia and East Africa. Particularly important is the fact that the middle Holocene was appreciably warmer than today. In Europe this phase has been called the Climatic Optimum (zones Boreal to Atlantic), and in North America it has been called the hypsithermal (also altithermal and xerothermic).

Macro-botanical remains

Like pollen, macrobotanical remains by themselves do not establish chronologies. Absolute dating of these remains does, however, provide a chronology of floral changes throughout the Holocene. Recent discoveries of the dung deposits of Pleistocene animals in dry caves and alcoves on the Colorado Plateau, including those of mammoth, bison, horse, sloth, extinct forms of mountain goats, and shrub oxen, have provided floristic assemblages from which temperature and moisture requirements for such assemblages can be deduced in order to develop paleoenvironmental reconstructions tied to an absolute chronology. Macrobotanical remains found in the digestive tracts of late Pleistocene animals frozen in the permafrost regions of Siberia and Alaska also have made it possible to build paleoenvironmental reconstructions tied to absolute chronologies.

From these reconstructions, one can see warming and drying trends in the terminal Pleistocene (± 11,500 BP). Cold-tolerant, water-loving plants (*e.g.*, birch and spruce) retreated to higher elevations or higher latitudes (as much as 2,500 metres in elevation) within less than 11,000 years.

Alluvial deposits

Detailed studies of late Pleistocene and Holocene alluvium, tied to carbon-14 chronology, have provided evidence of cyclic fluctuations in the aggradation and degradation of Holocene drainage systems. Although it is still too early in the analysis to state with certainty, it appears from the work of several investigators that there is a regional, or semicontinental cycle, of erosion and deposition that occurs every 250–300, 500–600, 1,000–1,300, and possibly 6,000 years within the Holocene.

Faunal change. According to an analysis of multiple carbon-dated sites conducted in 1984 by James I. Mead and David J. Meltzer, 75 percent of the larger animals (those of more than 40 kilograms live weight) that became extinct during the late Pleistocene did so by about 10,800 to 10,000 years ago. Whether the cause of this decimation of Pleistocene fauna was climatic or cultural has been debated ever since another American investigator, Paul S. Martin, proposed the overkill hypothesis (see above *Pleistocene fauna and flora*) in the 1960s. Whatever the case, it seems appropriate on paleontological grounds to designate the beginning of a new epoch—the Holocene—at approximately 10,000 years ago.

Floral and faunal reconstructions tied to the physical evidence of fluvial, alluvial, and lacustrine sediments and to a radiocarbon chronology reflect a warming and drying trend (as contrasted with the Pleistocene) during the last 10,000 years. The drying trend apparently reached its peak about 5,500 to 7,500 years ago (referred to as Antev's Altithermal) and has ranged between that peak and the cold, wet conditions of the early Holocene since that time.

Nonmarine Holocene sediments are usually discontinuous, making exact correlations difficult. An absolute chronology provided by radiocarbon dating permits temporal correlation, even if the deposits are discontinuous or physically different. Analysis of Holocene deposits requires chronostratigraphic correlations of discontinuous and dissimilar deposits to allow an interpretation of local, regional, continental, and global conditions.

Analysis of microfauna from paleontological and archaeological sites of the late Pleistocene and Holocene of North America has aided in paleoenvironmental reconstructions. Micromammals (rodents and insectivores), as well as amphibians and insects, are paleoecologically sensitive. Comparisons of modern habitat and range of species to late Pleistocene and Holocene assemblages and distributions reveal disharmonious associations (*i.e.*, the occurrence of presently allopatric taxa that are presumed to be ecologically incompatible), especially in late Pleistocene assemblages. Tentative conclusions from micromammals and other environmental indicators suggest that the late Pleistocene supported an environment in which there coexisted plants and animals that are today separated by hundreds to thousands of kilometres (or considerable elevation differences). Stated in another way, the late Pleistocene climate was more equable than that of the present day, one in which seasonal extremes in temperature and effective moisture were reduced. The evolution of a modern biotic community, as opposed to one of the late Pleistocene, appears to be the consequence of intricate biological and biophysical interactions among individual species. Some researchers have theorized that the environmental changes that led to the formation of new biotic communities at the end of the Pleistocene resulted in the extinction of many of the Pleistocene faunal forms.

Holocene climatic trends and chronology. In the mid-latitudes and the tropics, the end of the last glacial period was marked by a tremendous increase in rainfall. The increased precipitation toward the end of the Pleistocene was marked by a vast proliferation of pluvial lakes in the Great Basin of western North America, notably Lake Bonneville and Lake Lahontan (enormous ancestors of present-day Great Salt Lake and Pyramid Lake). Two peaks of lake levels were reached at about 12,000 ± 500 BP (the beginning of the Allerød Warm stage) and approximately 9000 ± 500 BP (the early Boreal Warm stage). At Lake Balaton (in Hungary) high terrace levels also mark the Allerød and early Boreal Warm stages. Lake Victoria (in East Africa) exhibits the identical twin oscillation in its terrace levels.

In equatorial regions the same evidence of high solar radiation and high rainfall at the end of the Pleistocene and during the early Holocene is apparent in the record of the Nile sediments. The Nile, like the other great rivers of Africa (notably, the Congo, Niger, and Sénégal), became very reduced, if not totally blocked, by silt and desert sand during the low-precipitation, arid phases of the Pleistocene. An erroneous correlation between glacial phases and pluvial phases in the tropics has been widely accepted in the past, although cold ocean water means less precipitation, not more. The pluvial phases correspond to the high solar radiation states, the last maximum being about 10,000 years ago. Thus tremendous increases of Nile discharge are determined, by radiocarbon dating, to have occurred around 12,000 and 9,000 years ago, separated and followed by alluviation, indicating reduced runoff in the headwaters.

The expansion of monsoonal rains during the early Holocene in the tropical latitudes permitted an extensive spread of moist savanna-type vegetation over the Sahara in North Africa and the Kalahari in South Africa and in broad areas of Brazil, India, and Australia. Most of these areas had been dry savanna or arid during the last glacial period. Signs of late Paleolithic and Neolithic people can be seen throughout the Sahara today, and art is representative of the life and hunting scenes of the time. Lake deposits have been dated as young as 5000–6000 BP. Lake Chad covered a vast area in the very late Pleistocene and up to 5000 BP. The Dead Sea throughout the early Holocene shows a record of sedimentation from humid headwaters; there was a Neolithic settlement at Jericho around 9000–10,000 BP.

Dispersal of early humans

In the high to mid-latitudes after the early Holocene, with its remnants of ice-age conditions (tundra passing to birch forests), there was a transition to the mid-Holocene, marked by a progressive change to pine forest and then oak, beech, or mixed forest. The mean annual temperature reached 2.5° C above that of today. Neolithic humans pressed forward across Europe and Asia. In the Canadian Arctic and in Manitoba the mean temperature passed 4° C above present averages. It was a "milk-and-honey" period for early humans over much of the world, and in Europe it paved the way for the cultured races of the Bronze Age. Navigators started using the seaways to trade between the eastern Mediterranean, the British Isles, and the Baltic.

In the mid-latitude continental interiors there was still evidence of hot summers, but the winters were becoming colder and partly drier. There was an expansion of steppe or prairie conditions and their associated fauna and flora. Many lake levels showed a fall.

In Europe there was also the beginning of widespread deforestation as Bronze Age human communities started to use charcoal for smelting and extended agriculture to tilling and planting. As a consequence, soil erosion began almost immediately, hillsides developed lynchets (terracettes), and "anthropogenic sediments" began to accumulate on the lower floodplains.

At the corresponding latitudes in the Southern Hemisphere (approximately 30° to 35° S), pollen analysis indicates increasing desiccation during the Subboreal stage, with a maximum dryness about 3200 BP.

In the subtropical regions of Mesopotamia and the Nile valley, people had learned to harness water. The stationary settlements, advanced agriculture, and mild climates favoured a great flowering of human culture. It is surmised that, when the normal floods began to fail, human ingenuity rose to the occasion, as attested to by the development of irrigation canals and machinery.

The Sub-Atlantic stage (2200–0 BP) is the last major physical division of the geologic record. Historically its beginning coincides with the rise of the Roman Empire in Europe, the flowering of the classical dynasties of China, the Ptolemies in Egypt, the Olmec of central Mexico and Guatemala, and the pre-Incan Chavín cultures of Peru.

The record of solar activity is disclosed by documentation of auroras in ancient Chinese court records and later by sunspot numbers. Both phenomena reflect solar activity in general, but correlation with weather records in the higher latitudes is complicated. Other indicators of climate, such as tree-ring analysis and palynology, were previously mentioned, but many documentary indications are also useful: the time of the cherry blossom festival in Kyōto, Japan, the freezing of lakes, the incidence of floods, blizzards, or droughts, the economics of harvests, salt evaporation production, some disease statistics, and so on. The water levels of closed basins such as the Caspian Sea and particularly the evaporite basin of the Kara-Bogaz-Gol Gulf reflect runoff to the Volga. The Dead Sea bears witness to eastern Mediterranean precipitation.

The main trends of the Sub-Atlantic are identifiable as follows:

Classical Roman Period. This time interval is marked by the Florida or Roman emergence in the eustatic record about the BC–AD boundary and succeeded by a transgression.

The solar record is not complete, but indications are for low activity. Records of rainfall kept by the astronomer Ptolemy (fl. AD 127–145) in Alexandria noted thunderstorm activity in every summer month, in comparison with the totally dry summer today, which suggests a slightly wetter overall pattern in this latitude.

In northern Europe and in other high latitudes, in contrast, the cool stage at the beginning of the 1st century AD may have been drier and more continental, as evidenced by dune building.

Warming and desiccation trend

Late Roman Period. After the 1st century AD there is evidence of a progressive rise in sea level. Roman buildings and peat layers were covered by the marine transgression in the Netherlands, southern England, and parts of the Mediterranean. At the same time, drying and warming trends were associated with alluviation of streams and general desiccation in southern Europe and North Africa. Similar alluviation occurred in the American Southwest. This warming and desiccation trend is evident also in the subtropics of the Southern Hemisphere. The solar activity record indicates a mean intensity comparable to that of the mid-20th century.

Post-Roman and Carolingian Period. This period extends roughly from AD 400 to 1000. The important invasions of western Europe by the Huns and the Goths may have been generated by deteriorating climatic conditions in central Asia. Radiocarbon dating and studies of the ancient Chinese literature have disclosed that, when the glaciers of central Asia were large, the meltwaters fed springs, rivers, and lakes on the edge of the desert, and human communities flourished. When there was a warm phase, the water supply failed and the deserts encroached. Thus, in central Asia (and the Tarim Basin) during the cool Roman Period, the Old Silk Road permitted a regular trade between Rome and China, where the Han dynasty was flourishing. During the Ch'in, Wei, and Chou dynasties this trade declined. During the T'ang dynasty (AD 618–907) there was a reopening of the trade routes, and likewise during the Yüan dynasty (AD 1206–1368). Marco Polo passed this way in AD 1271. Radiocarbon dates of the 8.6-metre-high lake level at Sogo Nur showed overflow conditions from AD 1300 to 1450, after which gradual, fluctuating, but progressive desiccation followed, and today the area is almost total desert.

In North America the Post-Roman-Carolingian Period was marked by warm temperatures in the northern parts, with mean paleotemperatures in central Canada about 1° C above the present. In the semiarid southwestern United States, the arroyos, washes, and ephemeral river valleys were filling slowly with alluvium (younger "Tsegi alluvium"), an indication that stream energy was generated by the summer flash floods. There were marginal retreats in almost all the mountain glacier regions of the world from the Alps to Patagonia.

In the tropical region of Central America there was the unexplained decline of the coastal Mayan people (Mexico and Guatemala) about the 10th century AD. The mountain Mayas continued to flourish, however, and it is possible that the high precipitation of this warming period introduced critical ecological limits to continued occupation of the (now) swampy coastal jungles.

The Viking-Norman Period. Approximately AD 1000–1250 the worldwide warm-up that culminated in the 10th century and has been called the early Medieval Warm Period or the "Little Climatic Optimum," continued for two more centuries, although there was a brief drop in mean solar activity in the period around 1030–70. During the 8th to 10th centuries the Vikings had extended as far afield as the Crimea and exploited coastal salt pans, the existence of which speak for seasonally high evaporation conditions and eustatic stability.

In the Arctic regions during the 10th, 11th, and 12th centuries there was widespread navigation by the Vikings. Partly in response to reduced sea-ice conditions and milder climates they were able to establish settlements in Iceland, southern Greenland (Erik the Red, *c.* 985), and in eastern North America (Vinland; Leif Eriksson, *c.* 1000). In Alaska, from tree-ring evidence, the mean temperature was 2° to 3° C warmer in the 11th century than today. Eskimos had settled in Ellesmere Island about AD 900. Records of sea ice off Iceland show negligible severity from 865–1200. Often the westerly storm tracks must have passed north of Europe altogether.

After a brief interval of cold winters in Japan, the cherry blossoms returned to early blooming in the 12th century. In the semiarid southwestern United States there appears to have been increased precipitation, leading to a spread of vegetation and agriculture. Pueblo campsites dated AD 1100–1200 are found on top of the youngest Tsegi Alluvium. The snow line in the Rocky Mountains was about 300 metres higher than today. Similar trends are recorded in the Southern Hemisphere, notably in Australia and Chile. The first immigration of Maori peoples into New Zealand probably occurred at this mild time.

The so-called little ice ages

"Medieval" Cool Period. This interval, extending roughly from AD 1250 to 1500, corresponds to the Paria Emergence in the eustatic record and has been called one of the "little ice ages" by certain authors. Solar activity records show a decline from 1250 to 1350, a brief rise from 1350 to 1380, and then a phenomenal low that lasted until 1500. Pollen records in northern Europe reveal rather consistently cool conditions, and smoothed mean temperature curves show a cumulative drop during this period. Stalactite studies in a karst cave in France showed a travertine growth peak (indicating cool, moist conditions) in 1450. In North America cool, moist conditions were widespread at first, becoming dry later. The arroyos and washes became filled with the Naha Alluvium, and the human population decreased markedly. There is pollen evidence of a temperature drop of about 1° C. This is the period of the "Great Drought." In the upper Mississippi valley the Indian cultures began a general decline, accompanied by a transfer from agriculture to hunting. It was similar in the western prairies, and it was this hunting culture that the first Spanish explorers encountered.

In the Canadian north the mean temperatures had dropped about two degrees below the previous high. In the Sierra Nevada, the Rockies, and Alaska there were glacial readvances, with evidence of a 2° C temperature drop. In the Arctic regions, the Eskimo economy underwent a marked change to adjust to these more extreme conditions, which amounted to about 5° or 6° C below the mean of the climatic optimum.

The Norse settlements in Greenland were abandoned altogether as the permafrost advanced. Pollen studies at Godthåb indicate a shift from a maritime climate to a cold, dry continental regime. The sea ice off Iceland reveals an extraordinary growth in severity, from zero coverage before the year 1200 to eight-week average cover in the 13th century, rising to 40 weeks in the 19th century, and dropping again to eight weeks in the 20th century. In Japan there were glacial readvances and a mean winter temperature drop of 3.5° C. Summers were marked by excessive rains and bad harvests.

The equatorial regions now began a marked desiccation, with a drop in level of all the great African lakes. The Nile suffered a decreased flow and alluviation.

South of the equator in the temperate belts there occurred a general return to cooler and wetter conditions that have continued (with oscillations) until the present time in southern Chile, Patagonia, southernmost Africa, southwestern Australia, and New Zealand.

Little Ice Age. Throughout most of what is commonly called the Little Ice Age (1500–1850) the mean solar activity was quite low, but positive fluctuations occurred around 1540–90 and 1770–1800. The main westerly storm belts shifted about 500 kilometres to the south, and for much of the time the northern latitudes came under cool continental conditions. Observed temperature series in Europe from Paris to Leningrad show large fluctuations until 1850.

Glacier advances are recorded in the Alps, in the Sierra Nevada, and in Alaska. Corresponding low sea levels are recorded by early tide gauge records in The Netherlands and Germany. Even in equatorial latitudes there are traces of mountain glacier advances (as in the Andes of Colombia).

Impact of technology

The Industrial Age (AD 1850–1950). The year 1850 started a brief warming trend that persisted for 100 years. It also approximates a critical turning point in climatic, sea level, glacial, and sedimentologic records. In many regions of central and southern Europe "anthropogenic" sediments (or cultural layers) began to appear in Neolithic times (early to mid-Holocene). Elsewhere in the world (*e.g.,* in North America, Australia, South Africa), however, this type of sedimentation began around the middle of the 19th century, depending on soil erosion stimulated by mechanized (disk) plowing, large-scale deforestation, and engineering activity. Thus independently of natural climatic change, the century 1850–1950, marked by anthropogenic aridification, proved to be a time of man-made deserts.

The Earth now is on a long-term cooling trend of the glacial-interglacial cycles and is likely to continue so for several thousand years, but there are numerous modulating influences, meteorologic, geologic, and man-made. These may accelerate or reverse the general trend, and from the point of view of human history they can well play a critical role. (R.W.F./L.D.A.)

BIBLIOGRAPHY

General works. Overviews are presented in CLAUDE C. ALBRITTON, JR., *The Abyss of Time, Changing Conceptions of the Earth's Antiquity After the Sixteenth Century* (1980); DON L. EICHER, *Geologic Time,* 2nd ed. (1976); WILLIAM B.N. BERRY, *Growth of a Prehistoric Time Scale: Based on Organic Evolution,* rev. ed. (1987); HENRY FAUL and CAROL FAUL, *It Began with a Stone: A History of Geology from the Stone Age to the Age of Plate Tectonics* (1983); ROBERT H. DOTT, JR., and ROGER LYMAN BATTEN, *Evolution of the Earth,* 4th ed. (1988); and REED WICANDER and JAMES S. MONROE, *Historical Geology: Evolution of the Earth and Life Through Time* (1989). A. HALLAM, *Great Geological Controversies,* 2nd ed. (1989), traces the development of the history of geology and of various, often contradictory, concepts. For the early recognition of the geologic cycle and the promulgation of uniformitarianism, see the classics themselves: JAMES HUTTON, *Theory of the Earth, With Proofs and Illustrations,* 2 vol. (1795, reissued 1972); JOHN PLAYFAIR, *Illustrations of the Huttonian Theory of the Earth* (1802, reprinted 1964); and CHARLES LYELL, *Principles of Geology,* 3 vol. (1830–33), available also in many later editions. Relevant developments in modern geologic sciences are discussed in DONALD R. PROTHERO, *Interpreting the Stratigraphic Record* (1989); RUTH E. MOORE, *Man, Time, and Fossils: The Story of Evolution,* 2nd rev. ed. (1961); MARTIN J.S. RUDWICK, *The Meaning of Fossils: Episodes in the History of Palaeontology,* 2nd rev. ed. (1976, reprinted 1985); W. LEE STOKES, *Essentials of Earth History: An Introduction to Historical Geology,* 4th ed. (1982); DON L. EICHER and A. LEE MCALESTER, *History of the Earth* (1980); and DON L. EICHER, A. LEE MCALESTER, and MARCIA L. ROTTMAN, *The History of the Earth's Crust* (1984). (G.D.J.)

Relative and absolute dating. Surveys of methodologies of ascertaining geologic time include FREDERICK E. ZEUNER, *Dating the Past: An Introduction to Geochronology,* 4th rev. ed. (1958, reissued 1972); PATRICK M. HURLEY, *How Old Is the Earth?* (1959, reprinted 1979); E.I. HAMILTON, *Applied Geochronology* (1965); G. BRENT DALRYMPLE and MARVIN A. LANPHERE, *Potassium-Argon Dating: Principles, Techniques, and Applications to Geochronology* (1969); and HENRY N. MICHAEL and ELIZABETH K. RALPH (eds.), *Dating Techniques for the Archaeologist* (1971). Stratigraphic geology as the basis of relative age measurement is discussed in W.B. HARLAND, A. GILBERT SMITH, and B. WILCOCK (eds.), *The Phanerozoic Time-Scale* (1964); JOHN W. HARBAUGH, *Stratigraphy and the Geologic Time Scale,* 2nd ed. (1974); ARTHUR HOLMES, *Holmes Principles of Physical Ge-*

ology, 3rd ed. rev. by DORIS L. HOLMES (1978); N.J. SNELLING (ed.), *The Chronology of the Geological Record* (1985); DEREK V. AGER, *The Nature of the Stratigraphical Record,* 2nd ed. (1981); and GILLES S. ODIN (ed.), *Numerical Dating in Stratigraphy,* 2 vol. (1982). The development of methods of absolute age measurement is examined in HENRY FAUL (ed.), *Nuclear Geology* (1954); HENRY FAUL, *Ages of Rocks, Planets, and Stars* (1966); the collected symposium papers published as *Radioactive Dating* (1963); ROBERT L. FLEISCHER, P. BUFORD PRICE, and ROBERT M. WALKER, *Nuclear Tracks in Solids: Principles and Applications* (1975); E. JÄGER and J.C. HUNZIKER (eds.), *Lectures in Isotope Geology* (1979); GUNTER FAURE, *Principles of Isotope Geology,* 2nd ed. (1986); and ROBERT BOWEN, *Isotopes in Earth Sciences* (1988). (T.E.Kr.)

Geologic history of the Earth. Reviews of Earth history based on information provided by the record of geologic processes include PRESTON CLOUD, *Oasis in Space: Earth History from the Beginning* (1988); L.R.M. COCKS (ed.), *The Evolving Earth* (1981); KENT C. CONDIE, *Plate Tectonics & Crustal Evolution,* 3rd ed. (1989); L.A. FRAKES, *Climates Throughout Geologic Time* (1979), a survey of all glaciations in Earth history; W.B. HARLAND *et al., A Geologic Time Scale* (1982), a well-documented account with many time charts; HAROLD L. LEVIN, *The Earth Through Time,* 3rd ed. (1988), with emphasis on sediments and fossils; R.J. O'CONNELL and W.S. FYFE (eds.), *Evolution of the Earth* (1981); CARL K. SEYFERT and LESLIE A. SIRKIN, *Earth History and Plate Tectonics: An Introduction to Historical Geology,* 2nd ed. (1979), with emphasis on the Phanerozoic Eon; STEVEN M. STANLEY, *Earth and Life Through Time,* 2nd ed. (1989); and BRIAN F. WINDLEY, *The Evolving Continents,* 2nd ed. (1984). Summaries of the geologic history of individual continents and countries include DEREK V. AGER, *The Geology of Europe: A Regional Approach* (1980); L. CAHEN and N.J. SNELLING, *The Geochronology and Evolution of Africa* (1984); ARTHUR ESCHER and W. STUART WATT (eds.), *Geology of Greenland* (1976); CHARLES S. HUTCHISON, *Geological Evolution of South-East Asia* (1989); D.V. NALIVKIN, *Geology of the U.S.S.R.* (1973; originally published in Russian, 1962); A.J. TANKARD *et al., Crustal Evolution of Southern Africa: 3.8 Billion Years of Earth History* (1982); and YANG ZUNYI, CHENG YUGI, and WANG HONGZHEN (TSUN-I YANG, YÜ-CH'I CHE'NG, and HUNG-CHEN WANG), *The Geology of China* (1986).

Precambrian Era. L.D. AYRES *et al., Evolution of Archean Supracrustal Sequences* (1985), presents good overviews of worldwide occurrences. KENT C. CONDIE, *Archean Greenstone Belts* (1981), offers a detailed synthesis. D.R. HUNTER (ed.), *Precambrian of the Southern Hemisphere* (1981), includes authoritative accounts of Australia, southern Africa, and South America. ALFRED KRÖNER and REINHARD GREILING (eds.), *Precambrian Tectonics Illustrated* (1984), reviews classic occurrences. S. MAHMOOD NAQVI and JOHN J.W. ROGERS, *Precambrian Geology of India* (1987), is a succinct review. Other surveys of Archean and Proterozoic stratigraphy and paleontology include those by E.G. NISBET, *The Young Earth: An Introduction to Archaean Geology* (1987); and by J. WILLIAM SCHOPF (ed.), *Earth's Earliest Biosphere: Its Origin and Evolution* (1983). (B.F.W.)

Paleozoic Era. Summaries of Paleozoic history and life are found in many of the general works cited in the beginning sections of this bibliography.

Cambrian Period: C.H. HOLLAND (ed.), *Cambrian of the New World* (1971), *Cambrian of the British Isles, Norden, and Spitsbergen* (1974), *Lower Palaeozoic of the Middle East, Eastern and Southern Africa, and Antarctica* (1981), and *Lower Palaeozoic of North-Western and West-Central Africa* (1985), are detailed surveys of Lower Paleozoic rocks. Correlation charts, explanatory notes on rocks and faunas, and extensive references are found in REINHARD WOLFART, *The Cambrian System in the Near and Middle East* (1983); J.H. SHERGOLD *et al., The Cambrian System in Australia, Antarctica, and New Zealand* (1985); W.T. CHANG, *The Cambrian System in Eastern Asia* (1988); KAISA MENS, JAN BERGSTRÖM, and KASIMIERA LENDZION, *The Cambrian System on the East European Platform* (1990); and VLADIMIR A. ASTASHKIN *et al., The Cambrian System on the Siberian Platform* (1991). MICHAEL E. TAYLOR (ed.), *Short Papers for the Second International Symposium on the Cambrian System, 1981* (1981), contains research reports from many parts of the world. Descriptions of Cambrian life-forms and those of succeeding geologic periods are found in RAYMOND C. MOORE *et al.* (eds.), *Treatise on Invertebrate Paleontology* (1953–), a comprehensive multivolume work; part A contains a discussion of Cambrian biostratigraphy. STEPHEN JAY GOULD, *Wonderful Life: The Burgess Shale and the Nature of History* (1989), analyzes a famous Cambrian biota and its significance in the history of life. (R.A.R.)

Ordovician Period: In addition to relevant sections of the general surveys cited above, discussions of the development of knowledge of Ordovician time correlations, climates, sea-level changes, and life are offered in R.J. ROSS, JR., "The Ordovician System, Progress and Problems," *Annual Review of Earth and Planetary Sciences,* 12:307–335 (1984); and in a collection of essays by DAVID L. BRUTON (ed.), *Aspects of the Ordovician System* (1984). (W.B.N.B.)

Silurian Period: C.H. HOLLAND and M.G. BASSETT (eds.), *A Global Standard for the Silurian System* (1989), provides working definitions for systemic, series, and stage boundaries at stratotypes in the United Kingdom and the Czech Republic, summaries of other reference localities, and reviews of the major index fossils. ANDERS MARTINSSON (ed.), *The Silurian-Devonian Boundary* (1977), applies the "golden spike" concept in defining a chronostratigraphic boundary. JAMES A. SECORD, *Controversy in Victorian Geology: The Cambrian-Silurian Dispute* (1986), presents a historical treatment of Murchison's development of the Silurian system and the conflict that arose therein. A.M. ZIEGLER *et al.,* "Silurian Continental Distributions, Paleogeography, Climatology, and Biogeography," *Tectonophysics,* 40(1–2):13–51 (1977), offers an especially thorough treatment of Silurian weather systems that remains valid despite minor revisions in paleogeography. A thorough review of issues related to tectonics, paleogeography, sea-level events, reef development, and sedimentary patterns is found in *Special Papers in Palaeontology,* no. 44 (1990), the whole issue being devoted to the papers of the first international symposium on the Silurian System that took place in early 1989. (M.E.Jo.)

Devonian Period: For the Devonian rocks, environment, and life-forms, see D.L. DINELEY, *Aspects of a Stratigraphic System: The Devonian* (1984); M.R. HOUSE, C.T. SCRUTTON, and M.G. BASSETT (eds.), *The Devonian System: A Palaeontological Association International Symposium* (1979); W.S. MCKERROW and C.R. SCOTESE (eds.), *Palaeozoic Palaeogeography and Biogeography* (1990); and N.J. MCMILLAN, A.F. EMBRY, and D.J. GLASS (eds.), *Devonian of the World: Proceedings of the Second International Symposium on the Devonian System,* 3 vol. (1988). (M.R.H.)

Carboniferous Period: Discussions of the period are included in general texts on historical geology, such as COLIN W. STEARN, ROBERT L. CARROLL, and THOMAS H. CLARK, *Geological Evolution of North America,* 3rd ed. (1979). The results of ongoing research in the field are presented at congresses and published as INTERNATIONAL CONGRESS ON CARBONIFEROUS STRATIGRAPHY AND GEOLOGY, *Compte rendu* (irregular), including materials in French, English, and German. The Decade of North American Geology project of the Geological Society of America includes works on Carboniferous geology in many of its series of publications. Another series, undertaken by the International Union of Geological Sciences, summarizes the Carboniferous geology of all continents: Carlos Martinez Diaz (ed.), *The Carboniferous of the World* (1983–). (W.L.Ma.)

Permian Period: In addition to the general texts cited above, specific references include CHARLES A. ROSS and JUNE R.P. ROSS, "Permian," in *Treatise on Invertebrate Paleontology,* part A, *Introduction—Fossilization (Taphonomy), Biogeography, and Biostratigraphy* (1979), pp. 291–350, and "Late Paleozoic Sea Levels and Depositional Sequences," in CHARLES A. ROSS and DREW HAMAN (eds.), *Timing and Depositional History of Eustatic Sequences: Constraints on Seismic Stratigraphy* (1987), pp. 137–149; CHARLES A. ROSS, "Paleozoic Evolution of Southern Margin of Permian Basin," *Geological Society of America Bulletin,* 97(5):536–554 (1986); L.L. SLOSS (ed.), *Sedimentary Cover, North American Craton, U.S.* (1988); and GARRY D. MCKENZIE (ed.), *Gondwana Six: Stratigraphy, Sedimentology, and Paleontology* (1987). (J.R.P.R./Ch.A.R.)

Mesozoic Era. For a survey of the era, see the general texts on historical geology cited at the beginning of this bibliography, especially CARL K. SEYFERT and LESLIE A. SIRKIN, *Earth History and Plate Tectonics: An Introduction to Historical Geology,* 2nd ed. (1979), which develops important aspects of Mesozoic plate divergence, mountain building, geosynclines, and life-forms and extinction events. JOHN D. COOPER, RICHARD H. MILLER, and JACQUELINE PATTERSON, *A Trip Through Time: Principles of Historical Geology,* 2nd ed. (1990), incorporates a narrative review of Earth history by eras. (L.Si.)

Triassic Period: A. LOGAN and L.V. HILLS (eds.), *The Permian and Triassic Systems and Their Mutual Boundary* (1973), includes papers of an international conference on the subject. Information on the fauna and flora of the period, as well as its paleogeography, is found in N.J. SILBERLING and E.T. TOZER, *Biostratigraphic Classification of the Marine Triassic in North America* (1968); E.T. TOZER, *A Standard for Triassic Time* (1967), "Triassic Time and Ammonoids: Problems and Proposals," *Canadian Journal of Earth Sciences,* 8:983–1031 (1971), and *The Trias and Its Ammonoids: The Evolution of a Time Scale* (1984); and A. HALLAM (ed.), *Patterns of Evolution as Illustrated by the Fossil Record* (1977). (A.Lo.)

Jurassic Period: Discussions of notable geologic features of the period include DAVID G. HOWELL, "Terranes," *Scientific American,* 253(5):116–126 (November 1985), a treatment of the origin and nature of microplates, with an analysis of western North American terranes that provides evidence of ancient oceanic crust; JOHN C. LORENZ, *Triassic-Jurassic Rift-Basin Sedimentology: History and Methods* (1988), an overview of basin research from the 19th century to the present, emphasizing the Hartford Basin in Connecticut; and JOHN MCPHEE, *Basin and Range* (1981), a discussion of the basin and range structure of North America. The results of significant stratigraphic research are presented in RALPH W. IMLAY, *Correlation of the Jurassic Formations of North America, Exclusive of Canada* (1952), a clarification of the biostratigraphic basis for the correlations. Life-forms of the period are addressed in the appropriate sections of the *Treatise on Invertebrate Paleontology* cited above; RAYMOND C. MOORE, CECIL G. LALICKER, and ALFRED G. FISCHER, *Invertebrate Fossils* (1952), dealing with the biology, morphology, ecology, classification, and biostratigraphy of the invertebrate phyla; ROBERT T. BAKKER, *The Dinosaur Heresies: New Theories Unlocking the Mystery of the Dinosaurs and Their Extinction* (1986); and JOHN H. OSTROM, "A New Look at Dinosaurs," *National Geographic,* 154(2):152–185 (August 1978), intended to dispel old myths about dinosaur evolution and to substitute an ecosystem-based extinction scenario for the formerly proposed climatic, volcanic, dietary, or extraterrestrial explanations. (L.Si.)

Cretaceous Period: Studies of the stratigraphy of the period include T. BIRKELUND *et al.*, "Cretaceous Stage Boundaries: Proposals," *Bulletin of the Geological Society of Denmark,* 33(1–2):3–20 (1984); J.M. HANCOCK and E.G. KAUFFMAN, "The Great Transgressions of the Late Cretaceous," *Journal of the Geological Society of London,* 136:175–186 (1979), comparing the transgressions in Europe, North America, and other areas; T. MATSUMOTO, "Inter-regional Correlation of Transgressions and Regressions in the Cretaceous Period," *Cretaceous Research,* 1(4):359–373 (1980), comparing the stable areas and relationships to tectonically active areas; and RICHARD A. REYMENT and PETER BENGTSON (compilers), *Events of the Mid-Cretaceous: Final Report on Results Obtained by IGCP Project* (1986), surveying mid-Cretaceous studies worldwide. Cretaceous rocks and environmental features are the subject of W.G.E. CALDWELL (ed.), *The Cretaceous System in the Western Interior of North America* (1975), a series of papers on a Cretaceous epicontinental sea; C.R. LLOYD, "The Mid-Cretaceous Earth: Paleogeography, Ocean Circulation and Temperature, Atmospheric Circulation," *Journal of Geology,* 90(4):393–413 (1982), providing quantitative and qualitative synthesis of climate, circulation, and temperatures; and WALTER KEGEL CHRISTENSEN, *Upper Cretaceous Belemnites from the Vomb Trough in Scania, Sweden* (1986), focusing on the biostratigraphy and biogeography of boreal Europe. Life-forms are examined by NORMAN F. SOHL, "Cretaceous Gastropods: Contrasts Between Tethys and the Temperate Provinces," *Journal of Paleontology,* 61(6):1085–1111 (1987), a comparative faunal history; C.F. KOCH and N.F. SOHL, "Preservational Effects in Paleoecological Studies: Cretaceous Mollusc Examples," *Paleobiology,* 9:26–34 (1983); and S.M. STANLEY, *Earth and Life Through Time,* 2nd ed. (1989). (C.F.K.)

Cenozoic Era. *Tertiary Period:* In addition to general geologic histories, other sources include STEPHEN JAY GOULD, *Time's Arrow, Time's Cycle: Myth and Metaphor in the Discovery of Geological Time* (1987); MAURICE GIGNOUX, *Stratigraphic Geology* (1955; originally published in French, 4th ed., 1950); M.J. HAMBREY and W.B. HARLAND (eds.), *Earth's Pre-Pleistocene Glacial Record* (1981); A.M. SPENCER (ed.), *Mesozoic-Cenozoic Orogenic Belts* (1974); and WILLIAM J. FRAZIER and DAVID R. SCHWIMMER, *Regional Stratigraphy of North America* (1987). Studies of the environment of this interval of Earth history include JOHN M. ARMENTROUT, MARK R. COLE, and HARRY TERBEST, JR., *Cenozoic Paleogeography of the Western United States* (1979); B.M. FUNNELL and W.R. RIEDEL (eds.), *The Micropalaeontology of Oceans* (1971); and KOTORA HATAI, *Tertiary Correlations and Climatic Changes in the Pacific* (1967). Flora and fauna of the period are studied in CHARLES B. BECK (ed.), *Origin and Early Evolution of Angiosperms* (1976); W.B. HARLAND *et al.* (eds.), *The Fossil Record* (1967); DONALD E. SAVAGE and DONALD E. RUSSELL, *Mammalian Paleofaunas of the World* (1983); and R.J.G. SAVAGE, *Mammal Evolution: An Illustrated Guide* (1986). (W.A.Be.)

Quaternary Period: General summaries of the physical and biological record of the Pleistocene Epoch are found in RICHARD F. FLINT, *Glacial and Quaternary Geology* (1971); TAGE NILSSON, *The Pleistocene: Geology and Life in the Quaternary Ice Age* (1983); and V. ŠIBRAVA, D.Q. BOWEN, and G.M. RICHMOND (eds.), *Quaternary Glaciations in the Northern Hemisphere* (1986). Regional surveys include W.F. RUDDIMAN and H.E. WRIGHT, JR. (eds.), *North America and Adjacent Oceans During the Last Deglaciation* (1987); STEPHEN C. PORTER and H.E. WRIGHT, JR. (eds.), *Late-Quaternary Environments of the United States,* 2 vol. (1983); N.J. SHACKLETON, R.G. WEST, and D.Q. BOWEN (eds.), *The Past Three Million Years: Evolution of Climatic Variability in the North Atlantic Region* (1988); and A.A. VELICHKO, H.E. WRIGHT, JR., and C.W. BARNOSKY (eds.), *Late Quaternary Environments of the Soviet Union,* trans. from Russian (1984). R.S. BRADLEY, *Quaternary Paleoclimatology: Methods of Paleoclimatic Reconstruction* (1985), is a well-written reference source. A historical account of the development of the glacial theory, the Quaternary climatic record, and hypotheses of climatic change is given in JOHN IMBRIE and KATHERINE PALMER IMBRIE, *Ice Ages: Solving the Mystery* (1979, reissued 1986). Megafaunal extinctions at the end of the Pleistocene are explored in PAUL S. MARTIN and RICHARD G. KLEIN (eds.), *Quaternary Extinctions: A Prehistoric Revolution* (1984). The biological record is emphasized in ANTONY J. SUTCLIFFE, *On the Track of Ice Age Mammals* (1985); and MICHAEL O. WOODBURNE (ed.), *Cenozoic Mammals of North America: Geochronology and Biostratigraphy* (1987). (W.H.J.)

A chronological survey of the Holocene Epoch is provided by ERNST ANTEVS, "Geologic-Climatic Dating in the West," *American Antiquity,* 20(4):317–335 (April 1955). O.K. DAVIS *et al.*, "The Pleistocene Dung Blanket of Bechan Cave, Utah," in H.H. GENOWAYS and M.R. DAWSON (eds.), *Contributions in Quaternary Vertebrate Paleontology* (1984), pp. 267–282; RUSSELL W. GRAHAM, HOLMES A. SEMKEN, JR., and MARY ANN GRAHAM (eds.), *Late Quaternary Mammalian Biogeography and Environments of the Great Plains and Prairies* (1987); and J.I. MEAD *et al.*, "Dung of Mammuthus in the Arid Southwest, North America," *Quaternary Research* 25(1):121–127 (1986), are paleoecological and paleontological studies. Human ecology is the subject of NEIL ROBERTS, *The Holocene: An Environmental History* (1989); and IAN TATTERSALL, ERIC DELSON, and JOHN VAN COUVERING (eds.), *Encyclopedia of Human Evolution and Prehistory* (1988). (L.D.A.)

Geography

Geography is the study of the surface of the Earth. The word is derived from the Greek words *geō* ("the Earth") and *graphein* ("to write"). The surface of the Earth is the interface of the atmosphere, lithosphere, hydrosphere, and biosphere. It provides the habitat, or environment, in which humans are able to live. This habitable zone has a number of special characteristics. One of the most important is the complex interaction among many physical, biologic, and human elements of the Earth, such as land surface, climate, water, soil, vegetation, agriculture, and urbanization. Another characteristic is the high variability of the environment from place to place—hot tropics to cold polar regions, dry deserts to humid equatorial forests, vast level plains to rugged mountains, and uninhabited ice caps to densely settled metropolitan areas. Yet another is the consistency with which significant patterns occur, which makes possible generalizations about distributions; obvious examples are measurements of temperature and rainfall, which are the most important climatic elements affecting farming and many other human activities. Geographic study is particularly concerned with location, with areal patterns, with the interrelationships of phenomena (especially of the relationship between human society and the land, as in ecology), with regionalization, and with ties among areas. Typical areas of inquiry include where people live; in what sort of patterns they are distributed over the Earth's surface; what factors of environment, resources, culture, and economic development account for this distribution; whether or not significant regions can be recognized by types of population, livelihood, and culture; and what types of movements and relations occur among places.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 10/33 and 10/36, and the *Index*.

This article is divided into the following sections:

ORIGIN AND DEVELOPMENT OF GEOGRAPHY

Early history. Human beings are inherently curious. They wonder how other lands and peoples differ from their own home and folk. The first recorded knowledge of such differences came mainly from the accounts of travelers. The 5th-century-BC Greek writer Herodotus was an outstanding early example of one who carefully recorded his personal observations made during many years of extensive travel. The Greek perception of the Earth was highly advanced: the philosophers Pythagoras and Aristotle believed it to be a sphere, and the Pythagorean Philolaus taught that it revolved around a central fire. In the 3rd century BC Eratosthenes of Cyrene, whose *Geographica* was the first work to have the word geography as its title, employed ingenious reasoning and measurements to produce a remarkably accurate calculation of the circumference of the Earth. He had observed that at Syene (modern Aswān, Egypt) at noon on the summer solstice the Sun was directly overhead, while at Alexandria it cast a shadow. By calculating the angle of the shadow and using the distance between Syene and Alexandria, Eratosthenes arrived at a figure of 250,000 stadia (or stades) for the Earth's circumference. The length of the stadium unit that he used is not known, but the range of modern estimates for the unit indicate that his figure exceeded the actual circumference at the Equator by about 15 percent or less. Eratosthenes' figure, however, subsequently was rejected by classical geographers, such as Ptolemy—who calculated, erroneously, that the Earth was much smaller—and centuries would pass before a more accurate calculation would be made and generally accepted.

The ancient Greeks

In his 17-volume work written at about the time of Christ, the Greek geographer and historian Strabo provided the most detailed summary and review of the classical knowledge of geography. The first two books were devoted to a wide-ranging discussion of the aims and methods of geography and to a review of earlier writings; many early works of Greek or Roman authors have disappeared or have survived only in fragments, and they are known today only through Strabo's critical comments in these books. The other 15 books written by Strabo provided regional descriptions. The great contribution of the 2nd-century-AD astronomer and geographer Ptolemy was the concept of the tabulation of latitude and longitude of places; these tabulations could give precision to locations, but Ptolemy's data again contained errors that were to plague geographers for centuries.

With the breakup of the Roman Empire in the West, most of the geographic knowledge of the Greeks gradually was lost in Europe, but during the 11th and 12th centuries it was preserved, revised, and enlarged by Arab geographers. Geographic study in Europe was stimulated anew by contact with Muslim learning during the Crusades, although in their reacquaintance with Greek ideas—particularly those of Ptolemy—European thinkers generally ignored the additions and corrections of the Arabs. Thus, the errors of Ptolemy were perpetuated in the West until the voyages of the 15th and 16th centuries started bringing back to Europe detailed and more accurate information on the rest of the world.

The Arabs

An important figure of the new learning was the German scholar Bernhardus Varenius (Bernhard Varen), whose *Geographia generalis* (1650; *General Geography*) was revised numerous times and remained a standard reference work for a century or more. Unlike many earlier writers, Varenius included ideas based on direct observations and original measurements. In the century before Varenius, the Flemish cartographer Abraham Ortelius prepared a world map in sections and bound them together in book form in his *Theatrum orbis terrarum* (1570; *Epitome of the Theatre of the World*), the first atlas. The first use of the term atlas, however, was by Ortelius' contemporary Gerardus Mercator (Gerhard de Cremer) and is said to be derived from the representation of Atlas supporting the heavens that formed a frontispiece to early atlases. Mercator, who also came from Flanders, was the leading cartographer of the 16th century.

The four generations of the Cassini family of astronomers and surveyors in France were preeminent in developing methods for accurately surveying the land surface. In work extending from the late 17th to the late 18th century, the Cassinis made the first detailed topographic survey of a large country, and this was used as the basis for a national atlas of France published in 1791. In the 18th century James Cook set new standards in accuracy and skill in navigation. Furthermore, his voyages had scientific missions. On his famous second voyage (1772–75), which circumnavigated the globe at high southern latitudes, he was accompanied by Johann and Georg Forster, the father and son who made botanical collections and climatological observations. Georg Forster later influenced Alexander von Humboldt to study geography.

Contributions of Humboldt

Like many before him Humboldt was determined to see for himself what other parts of the world were like, but he was distinguished by the range of his curiosity, by the careful preparation for his trips, by the range and precision of his observations and measurements, by his mental acuity in suggesting and evaluating possible interrelationships, and by his utilization of carefully recorded notes and observations over many years for reflective analysis and publication. He purchased, learned to use, and made extensive measurements with the measuring instruments of his time. During his travels in South and Central America (1799–1804) Humboldt located places with accurate latitudes and reasonably close longitudes. Through his detailed observations in the Andes he was able to provide the first systematic description of the interrelations of altitude, temperature, vegetation, and agriculture in low-latitude mountains and to provide a clear picture of vertical zonation. He plotted his data on maps and coined the term isotherm for a line joining points with the same temperatures. In his regional monograph on the economic geography of New Spain (Mexico), Humboldt presented data on population, production, trade, utilization of resources, and their interconnections.

Emergence of the modern discipline. Humboldt thus laid the groundwork for modern geography, with its emphasis on direct field observation and accurate measurements as the basis for generalizations. Three institutional innovations in the 19th century also played important roles in the development of the modern discipline: a new type of university, the rise of geographic societies, and organized government surveys of natural conditions and resources. In 1809 Humboldt's brother Wilhelm, with the support of Frederick William III of Prussia, founded the University of Berlin (now the Humboldt University of Berlin), without requirements to either faculty or students to adhere to a particular creed. Although universities had existed since the Middle Ages, many had come to be viewed primarily as defenders of Christianity rather than as centres of free inquiry. The flowering of universities gave rise to specialists in many academic disciplines, including geography, who besides teaching carried out programs of original research that created great bodies of new knowledge.

In his *Kritik der reinen Vernunft* (1781; *Critique of Pure Reason*) the 18th-century German philosopher Immanuel Kant had already provided a reasoned statement of the place of geography among the fields of learning, noting that geography dealt with phenomena associated in space just as history dealt with events occurring together in time. Both Kant and Alexander von Humboldt lectured on physical geography, the latter at the University of Berlin. Carl Ritter, a contemporary of Humboldt, also taught at the university, occupying the first chair of geography ever established at a modern university.

Founding of geographic societies

As the number of teachers and specialists grew, they created professional associations. Geographic societies began to play an important role in the diffusion of geographic knowledge. The Société de Géographie de Paris was founded in Paris in 1821, the Gesellschaft für Erdkunde zu Berlin in Berlin in 1828, the Royal Geographical Society in London in 1830, the Imperial Russian Geographical Society in St. Petersburg (modern Leningrad) in 1845, and the American Geographical Society in New York City in 1851. These associations and others provided a forum for reporting on new geographic knowledge by sponsoring lectures and discussions and by publishing periodicals, monographs, and maps.

The rise of governmental surveys of resources, natural features, and possibilities for settlement was particularly important in the large continental expanses of the United States and the Russian Empire. In the United States Thomas Jefferson, who had a keen sense of geography, dispatched the Lewis and Clark Expedition (1804–06) to explore the newly acquired lands of the Louisiana Purchase and beyond. In the second half of the 19th century a number of expeditions to the West surveyed routes for a transcontinental railroad, explored possibilities for settlement, determined the extent of mineral resources, and mapped the basic characteristics of the terrain. The Russian tsar Peter I the Great recognized the importance of exact geographic knowledge in the landward expansion of his country and thus supported expeditions and geographic publications. In 1727 Ivan Kirilov produced the first systematic geographic and economic description of Russia and in 1734 published the first atlas of the Russian Empire. Many geographic expeditions were sponsored by the government and by geographic societies. The establishment of networks for systematic geographic observation aided mapping of many physical phenomena. At the suggestion of Humboldt, a network of weather stations was established throughout Russia, and Humboldt later was able to utilize data from these stations in his studies of continental climates. While he was in charge of the U.S. Navy's Depot of Charts and Instruments (later the U.S. Naval Observatory), the pioneer hydrographer Matthew Fontaine Maury distributed special logbooks to sea captains for the recording of information on winds and currents. From these observations he was able to prepare maps of the Earth's major wind patterns, which enabled vessels to shorten considerably the time of voyages. In 1855 Maury published *The Physical Geography of the Sea,* the first work of modern oceanography.

Geography as an academic discipline—*i.e.,* as a field of advanced study and research in universities—became well-established in Germany in the 1870s. In 1874 the Prussian government decreed that a chair of geography (to be occupied by a professor) would be established in each of the Prussian universities, and by 1880 10 such professors had been appointed. Other universities in Germany, as well as in France and other European countries, quickly followed suit. Among the leaders during this period of expansion and definition of the field was the German geographer and geologist Ferdinand Paul Wilhelm, Freiherr von Richthofen, who wrote a monumental five-volume study of Chinese geography and was influential in the development of geographic methodology in Germany and elsewhere. Another German, Friedrich Ratzel, wrote pioneering works in the fields of human and political geography.

Rise of opposing schools

The rise of a considerable body of geographers with advanced academic training in the discipline led to the emergence of differing beliefs on the nature of geography. One school thought of geography as the study of the interrelationships of phenomena that occur together in space, specifically of the relationships of man and nature. In one form this was conceived as the influence that natural conditions exercise on humans—a sort of environmental determinism. The reverse of this was man's effect on nature, first examined by the American diplomat and scholar George Perkins Marsh in *Man and Nature, or Physical Geography as Modified by Human Action* (1864). A second school, founded on ideas of Richthofen and elaborated by the German Alfred Hettner, conceived of geography as areal differentiation. This concept was further expanded by the American geographer Richard Hartshorne in *The Nature of Geography* (1939). Yet a third school considered the proper object of geographic inquiry to be the visual landscape (either alone or in association with nonvisual elements). In this tradition an attempt was made to reconstruct the "original landscape" (*Urlandschaft*) of Earth before the advent of humans; this landscape was distinct from the cultural landscape, transformed by human agency. Other geographers interpreted landscapes mainly on the basis of vegetation types of the world and associated phenomena. Although differing in viewpoint and emphasis, each of these schools was concerned with the variations of human, physical, and biologic phenomena over space on the Earth's surface and the interrelationships or associations of these phenomena in specific places or regions.

The rise of modern geography in France was led by Paul Vidal de La Blache. Vidal de La Blache defined the field of regional geography, emphasizing the study of small, relatively homogeneous areas that he called *pays*. He also concentrated on the interrelationship of man and environment, which he termed "possibilism": within limits, nature offers possibilities to humans, who in utilizing these possibilities—expressed by such means as culture and technology—modify nature. In 1898 Vidal de La Blache became the first occupant of the chair in geography at the Sorbonne whose academic training was in geography. There he trained a large and able group of students, who themselves were named to professorships at other French universities and who wrote a distinguished series of works on the regions of France. Vidal de La Blache had planned a mammoth series of regional geographies to cover the entire world but died before the plan could be realized; the 23-volume *Géographie universelle* (1927–48), completed mostly under the direction of his student Lucien Gallois, has remained the most successful world regional geography ever published.

In Great Britain a report by the Royal Geographical Society, in which British scholarship in geography was compared unfavourably with that in the countries of continental Europe, prompted the appointment of geographers at the universities of Oxford and Cambridge in 1887 and 1888, respectively. Halford (later Sir Halford) J. Mackinder, who became the chair at Oxford, had a grand sweep of both history and geography. His principal work, *Britain and the British Seas* (1902; 2nd ed. 1930), included an imaginative and broad interpretation of the rise of Britain and of British sea power in relation to the maritime resources and protection from invasion provided by the surrounding seas. He later developed the concept of a landlocked heartland in Eurasia inaccessible to sea power and a rimland easily accessible from the ocean and of the relations between the two regions. This concept was further developed after World War I in his *Democratic Ideals and Reality* (1919); and still later it found ultimate expression in the creation of the North Atlantic Treaty Organization and the Warsaw Pact political and military blocs. Others who contributed to the rise of geography in Britain include Hugh Robert Mill, who studied the part played by water in the world economy; George G. Chisholm, whose *Handbook of Commercial Geography* (1889; many subsequent editions) provided a meticulous, detailed worldwide survey of commercial products and conditions of their production and trade; and A.J. Herbertson, who succeeded Mackinder at Oxford and who proposed a framework of natural regions for the study of world geography.

Work of Mackinder

In prerevolutionary Russia four individuals played key roles in the development of geography in the late 19th century: P.P. Semyonov, who wrote a seminal five-volume study of Russian geography; A.I. Voyeykov, who investigated climate, particularly in relation to agriculture; V.V. Dokuchayev, who made pioneering studies in soil geography; and D.N. Anuchin, who in 1885 became the head of the new department of geography at the University of Moscow and who trained many of the geographers who later occupied key positions in the Soviet Union.

Geography in the 20th century. At the end of the 19th century geography was a subject of original research and teaching mainly in the universities of Germany, France, and nearby continental countries and, to a lesser extent, in the major universities of Great Britain. The flowering of geography in these countries, coupled with the excitement of the great surveys of the American West, had a strong impact on geographic studies at the universities in the United States. During the first half of the 20th century geography became well-established in the United States, as well as in Great Britain. A key figure in the United States was William Morris Davis, who contributed to physical geography—especially to the branch called geomorphology—by developing the concept of the erosion cycle. Many of his students became distinguished contributors to geography. An important event in the rise of academic geography in the United States was the creation in 1903 at the University of Chicago of a department of geography devoted to graduate instruction and research. It trained many of the geographers who were among the earliest professors and leaders in the development of the modern discipline in American universities. Also important was the Department of Geography and Industry at the University of Pennsylvania, Philadelphia, where J. Russell Smith wrote the influential *Industrial and Commercial Geography* (1913), a highly readable regional geography of North America that in many revisions remained a major textbook for half a century. The teaching of geography was cultivated in many teachers' colleges, among which Michigan State Normal College (now Eastern Michigan University), in Ypsilanti, was outstanding.

Geography in the United States

After World War II, particularly in the 1950s and '60s, geographic studies underwent considerable expansion in the United States and Great Britain and spread widely in the universities of Canada, Australia, and Japan. In eastern Europe, particularly in what was then the Soviet Union, Poland, and Hungary, important research institutes were established within the framework of academies of sciences. Scandinavia became a centre of study of the spatial diffusion of innovation: particularly influential was the work of Torsten Hägerstrand, who studied the impact of the growing use of automobiles and telephones in Sweden. Contrasts of geographic work among countries resulted from varied national traditions, differing needs, and unequal receptivity to change.

International ties among geographers were greatly strengthened by three factors: active programs sponsored by the International Geographical Union, which has members from some 90 countries who participate in international congresses and contribute to the work of commissions; the evolving role of English as a common language of international communication and conferences and its supplementary use in geographic periodicals in other languages or as a basic language of scientific publications in many non-Anglophone countries; and the speed and ease of international travel, facilitating frequent personal contacts.

During the 20th century the discipline has evolved rapidly, with many new concepts and methodologies. At first geomorphology was the exciting field, with the generalizing power of Davis' theories. Until mid-century the focus was particularly on regional geography and the great diversity of the lands and peoples of the world. After World War II regional geographers often were associated with the implementation of economic programs in less-developed areas. In the 1960s attention turned to the development of quantitative methods and the building of models of both physical and human systems. In the late 1960s and the 1970s active environmental concerns came once again to the fore after a period of relative neglect that had resulted from a reaction to an earlier, crude environmental determinism. In the 1970s and '80s the rise of quantitative models based on large bodies of data from censuses and other surveys came to be perceived by some geographers as dealing too much with abstract space and not enough with terrestrial place. Calls were made for a behavioral approach involving individual perceptions and choices and for a more humanistic geography. Still others called for more radical approaches. But throughout these changes ran a common thread of intense interest in human beings and their societies, in the physical and biologic environment, in areal patterns of occurrence on the surface of the Earth, and in associations and interrelationships in specific places and regions, whether viewed from an ecological or systems perspective. Better statistical sources became available, especially in many new national censuses. Aerial photography offered a new tool, but even more exciting and powerful have been the possibilities of remote sensing from artificial satellites and of using computers to handle masses of data and their analysis.

Impact of human occupation

The very object of geographic study—the surface of the Earth—has changed rapidly, particularly under the impact of humans, in the second half of the 20th century. Geographers, in common with scientists and scholars in many fields, increasingly have become concerned with a number of problems: desertification, whether caused by recurring droughts or by human actions; the hasty clearing of equatorial forests, which upset delicate biological balances; the threat of natural disasters of many kinds and also of man-made accidents, particularly nuclear ones; environmental pollution, as with acid rain in North America and Europe or air pollution in cities; the high rates of population increase that pose special problems of survival in some developing countries with limited resources; the persisting problems of enormous regional variations in resource use, productivity, wealth, and welfare; and the spectre of hunger and starvation in fragile environments, often exacerbated by economic and political problems. Among the potential developments have been exploitation of the mineral and other resources of the oceans and ocean floors, the use of biogenetic engineering to increase agricultural productivity and to solve problems created by the pests that inhibit expansion of farming in many regions of the world, and the perfection of superconductivity to overcome the inefficient transfer of electrical energy. All of these problems and potentials include geographic attributes—since they involve natural and human factors in a spatial setting—and they present continuing frontiers for research and study.

GEOGRAPHIC METHODS

Map location and measurement. The map is the distinctive data bank of the geographer. Since geography deals particularly with locations, distributions, areal associations, and interrelationships of phenomena in space, accurate observation and measurement of the surface of the Earth and the recording and displaying of location on maps are of prime importance.

Latitude and longitude

Latitude and longitude are commonly utilized for plotting locations on the surface of the globe. Fairly accurate measurements of latitude were made in antiquity by Greek scholars. Measurements of longitude remained rough, however, because of the difficulty in measuring differences in solar time (the Sun "moves" westward at the mean rate of one degree each four minutes). The perfection of the chronometer solved this problem, but for long each country had its own system for numbering the meridians. Finally, by an international agreement reached in 1884, an imaginary line from pole to pole through Greenwich, near London, was recognized as the prime meridian (*i.e.,* 0° longitude). Measurement of direction, or bearing, was aided considerably by use of the magnetic compass, but as Christopher Columbus noted in crossing the Atlantic, the direction in which the compass pointed varied with longitude.

The measurement of distances overland could be counted in days of journey on foot, by camel, by horse, or by other means. More accurate measurements of short distances were obtained by using a chain, and the chain as a unit of length (66 feet) is still a traditional surveying measure in English-speaking countries. Later, the chain itself was replaced by a steel tape, and still later electronic instruments came into use. A practical measurement of distances at sea was developed in the 16th century: a log was thrown overboard and the amount of time it took the stationary log to play out a certain distance on a line marked off with knots was measured. Navigation by means of satellites is now available, but a ship's speed is still measured in knots and records are kept in a logbook. After the adoption of the metre as a standard unit in France in the late 18th century, it gradually replaced older local and national measures of distance over much of the world during the 19th and 20th centuries.

Maps of small areas—topographic maps, for example—can be made by a method called triangulation. A baseline is measured with chains or other devices, and by using this base as one side of a triangle the other sides are calculated from the angles at the two ends of the baseline. Angles can be measured more easily and accurately than distances, and from the points on the corners of the original triangle, a network of points joined by triangles can be established. Triangulation was known to the ancient Egyptians and Greeks; with improved instruments, especially the theodolite, this method was utilized in the great national surveys of Europe and America from the 18th to the 20th century. How to represent the entire, spherical Earth or large areas of it on maps remained a problem. In 1492 the German navigator and geographer Martin Behaim completed the construction of a terrestrial globe. Ships following straight lines on flat maps, however, did not arrive at expected points. Mercator devised a map projection—which became known as the Mercator projection—on which ships following straight lines would arrive at the plotted points. Although the projection was excellent for navigation, it was poor for many geographic comparisons, since the size of areas in the higher latitudes becomes grossly exaggerated; Greenland, for example, appears to be larger than South America, though in fact it is less than one-eighth as large. The Earth cannot be shown accurately in all respects on a flat piece of paper since either bearing (angle), distance, or scale must be distorted. Modern geographers use maps drawn in what is called the equal-area projection, but even this projection distorts shapes or distances, particularly toward the edges of the map.

Map projections

With the increasing specialization of knowledge, measuring the shape of the Earth developed into the discipline of geodesy, plotting land positions for detailed maps became the province of surveying, and constructing numerous types of maps with appropriate projections grew into the field of cartography. Maps have remained as the basic tools in geography for plotting and analyzing a vast range of physical, biologic, historical, economic, political, and social data. Geography and cartography are closely associated: many geographers are cartographers and a significant number of cartographers are geographers, and cartographic training is widely available in departments of geography in universities and colleges.

Aerial photography and remote sensing. During the 20th century immense strides have been made in observing features on the Earth's surface, first by the development of aerial photography and later by satellite imagery. Aerial photography was first used extensively during World War I for reconnaissance. A new profession of photographic interpretation evolved in which the identification of both natural and man-made features became a special skill. After the war aerial photographs quickly proved their usefulness in mapping landforms, land uses, types of forest, vegetation, and the limits of the built-up areas of cities; in locating archaeological sites; and, in the United States, in measuring the size of farm fields and crop acreages in agricultural programs of the federal government and as the basis for county soil surveys. The perfection of stereoscopic plotting from overlapping aerial photographs greatly improved contour mapping in the production of topographic maps. The development of side-looking airborne radar made possible rapid surveys of large tracts of land, particularly in areas of economically underdeveloped countries that were not easily accessible by other means. Colour photographic film and infrared imagery also increased the potential for measurement of many aspects of land use or of physical processes.

Use of satellites

Even more revolutionary was the rapid evolution from the late 1950s of remote sensing using artificial satellites. Simultaneous, or synoptic, worldwide patterns could be viewed for the first time. The first weather satellite, called Nimbus, was launched in 1964. The advances in meteorology provide a particularly vivid example. Although humans had been observing the weather for countless ages and all major countries had dense networks of weather stations, satellites finally made it possible to clearly recognize many cloud patterns, such as the huge spiral cloud formations of major weather systems. By using weather satellites in combination with communication satellites, cloud and rainfall patterns could be presented almost instantaneously to worldwide television audiences. The rapid advances in space technology and multiband remote sensing allowed for a wide range of physical and man-made features to be observed, mapped, and studied with a comprehensiveness not previously possible. Data from geodetic satellites gave a much more exact determination of the shape of the Earth and revealed many irregularities not previously recognized. With the advent of navigation satellites locations could be determined with much greater precision. When astronauts traveled to the Moon people saw for the first time the whole terrestrial globe as a unity (though, of course, only half could be seen at one time).

Mathematical and statistical analysis. Although the use of mathematics in geography is ancient, the extensive and fundamental utilization of quantitative and statistical methods in modern geography arose mainly in the second half of the 20th century. The methodologies are similar to or identical with those used in the physical, biological, and social sciences, except that special attention is given to the problems of place-specific or area-specific data, in which the recognition and description of spatial patterns require special techniques. Model building, probability theory, and simulation techniques have proved to be of special value. Torsten Hägerstrand, for example, utilized an excellent data base in his native Sweden to make his seminal studies of the geographic diffusion of innovations. The development of electronic computers greatly expanded the power to analyze geographic problems. Computers can store immense amounts of data, and they greatly accelerate the speed at which complicated statistical and mathematical problems can be solved. They have become particularly valuable for their ability to handle such programs as those that measure spatial contiguity, spatial diffusion through time, and locational patterns, as well as for their use in network analysis, node accessibility indexes, map projections, and the display of cartographic data.

Importance of direct observation

There is a long-standing tradition in geography of placing special value on direct field observation and mapping. Much geographic knowledge is based on observations, if not by a research investigator directly then by others who write about observations revealed in oral interviews, in census enumerations, or in the interpretation of remote sensing images, aerial photographs, or maps. The questions others ask may not be formulated in such a way as to reveal the locational characteristics and spatial patterns so important to the geographer. Census information, for example, often fails to be broken into units small enough to permit satisfactory areal analysis. Thus, geographers often need to gather their own data.

FIELDS OF MODERN GEOGRAPHY

Geography is divided into systematic fields and regional specializations, which can be grouped under three main headings: physical geography, human geography, and regional geography.

Physical geography. The principal activities of the physical geographer—observing, measuring, and describing the surface of the Earth—are those aspects of the larger discipline of geography that are the most recognizable to the layperson. Even so, the growing complexity of geographic inquiry has resulted in increased specialization within the field. The principal branches of physical geography are geomorphology, climatology, biogeography, and soil geography. As human activity has become more able to affect the landscape and ecology of the world, two more branches have emerged: resource management and environmental studies.

Geomorphology. Geomorphology is the study of the forms of the land's surface and of the processes that mold them. As the branch has developed, landforms increasingly have been described with quantitative measures. The processes of weathering, erosion, transport, and deposition are analyzed. Fluvial, glacial and periglacial, coastal, and eolian processes are recognized and are often the subjects of specialized study. Slopes have received detailed consideration. In the branch called climatic geomorphology, the special landforms developed in deserts, in humid tropical conditions, and in the glacial and periglacial conditions of polar areas (or in relicts from the ice ages in other areas) are examined. Landforms are also changed by human activity, either directly (such as by constructing cities, excavating mines, or modifying hydrology by the building of dams and reservoirs) or indirectly by modifying vegetation (clearing forests, cultivation, and overgrazing), which then may accelerate geomorphologic processes by increasing runoff and causing severe gullying and soil erosion.

Applied geomorphology

Applied geomorphology increasingly is being utilized in environmental management. Engineering measures sometimes have unanticipated environmental consequences that may intensify rather than reduce the problems that they mean to correct, as in some projects to control coastal erosion and flooding; understanding the complexity and interrelatedness of environmental conditions is the object of much of the newer research in geomorphology.

Climatology. Climatology is concerned with the prevailing state of the atmosphere, including average climatic values, seasonal and diurnal rhythms, extreme values, the frequency of values within stated ranges, weather types and their characteristics, and the explanation and distribution of both climatic elements and general climatic types. Of particular geographic interest are the interrelationships of climatic elements and types with other physical and biologic features and with human activities. Examples include the effect of insolation at the Earth's surface on temperature, atmospheric pressure, winds, and precipitation; the effect of the distribution of land and water on climatic elements; and the effect of climate on vegetation, soils, and agriculture. Because climates show extreme variations over the face of the globe—from equatorial regions to the poles and from oceanic areas to continental interiors—and are associated with many other features, they have high geographic interest. Many human activities also are affected by weather, from heating and cooling requirements in buildings to transportation (especially by air). Extreme weather conditions can cause great damage either directly by such phenomena as ice, hail, or wind (cyclones or typhoons) or indirectly through flooding caused by heavy or prolonged rain, especially if associated with spring thaws and melting snow. Unseasonable frosts may damage specialty fruit areas.

Three branches

Biogeography. Biogeography encompasses three different paths of geographic inquiry, although it is not uncommon for the distinctions among them to become blurred. As phytogeography or zoogeography, biogeography refers to the spatial distribution, past and present, of individual categories of plants or animals and of attempts to explain this distribution. Biogeography may refer to the interrelationships of plants and animals with one another and with environmental conditions, particularly within regional types of continental scale (called biomes) or within smaller areas of individual plant and animal communities, as in ecology. It is also sometimes used to refer to the study of the effects of human activity on plants and animals in the biosphere. Particularly fruitful to the growth of the field were the studies in the ecology of high mountains by the German geographer Carl Troll and his students.

Soil geography. Soil geography is concerned with the areal distribution of soil types over the land, the principal factors in formation of different types of soil, the plant and animal world as the source of the organic matter in the soil, horizons in a vertical profile of the soil, and the moisture, drainage, and erosional conditions that underlie the chain, or catena, of soil differences along the horizontal profile of a slope, from well-drained relatively flat upper surfaces down steeper slopes to low-lying relatively flat surfaces. Of special concern are the role of soils in agricultural systems and the impact of human activity on the improvement, erosion, degradation, or reclamation of soils.

Resource management and environmental studies. The broad fields of resource management and of environmental studies are of particular interest to geographers, for they involve both physical and biologic systems on the one hand and human systems on the other, and they typically have specific spatial patterns on the surface of the Earth.

Emphases of the two branches

Resource management tends to emphasize human direction in the utilization of natural resources for the benefit of humans, usually on a sustained-yield or long-range basis, as, for example, in the development of the water resources of a stream for multiple purposes (power, irrigation, and recreation). Environmental studies tend to focus on the plant and animal worlds and on the threats to them posed by human activities; on the degradation of the atmosphere, hydrosphere, and lithosphere by pollution of many types; or on the interaction of these two aspects, as by acid rain created from the production of energy from hydrocarbons or by the potential depletion of the ozone layer in the upper atmosphere through the use of chlorofluorocarbons. In all of these studies geographers generally take into account alternative technologies, relative costs, impacts on other systems, perceptions of dangers and gains, policy alternatives, and the spatial range of the benefit or problem.

Specific man-made environmental problems that have been studied geographically include accelerated soil erosion through crop intertilling, overgrazing, and deforestation; the human impact on stream regimen and quality; the environmental impact of modern agricultural technologies through the increased use of fertilizers, pesticides, and herbicides; the impoverishment of natural species directly or indirectly through human actions; and the environmental impact of urbanization.

Early research in the utilization of water resources, for example, was concerned mainly with engineering studies for the construction of large dams, but later studies were expanded to include alternative programs for conserving water in irrigation or for conserving energy use. The initial emphasis in flood control was on building levees—which concentrate flood flows in restricted channels—permitting unrestricted construction in natural floodplains; later studies have included analyzing land-use regulation in floodplains, improving flood-warning systems, waterproofing buildings in areas subject to flood, and other measures in an effort to reduce flood damage and losses at far less cost and with greater effectiveness. The role of water quality in maintaining environmental systems also has received increased emphasis. The American geographer Gilbert F. White has played a leading international role in studies of the management of resources and of environmental hazards.

Human geography. One of the central problems in human geography is to explain the distribution and characteristics of people—this is the province of population geography. But this distribution can be understood only if attention is paid to how people satisfy their needs and make a living, the field of economic geography; to their cultural and social values, tools, and organization, which are the fields of cultural and social geography; to their concentrations in cities and metropolitan areas, the object of urban geography; to their political organization, examined by political geography; to their health and to the diseases that affect them, the field of medical geography; and to the evolution of their present patterns, the subject of historical geography.

Focus on distribution

Population geography. Population geography examines particularly the distribution of population in relation to its various characteristics, such as growth, number, density, age, sex, fertility, mortality, natural increase, and occupations; division into rural and urban, ethnic, linguistic, or religious groupings; and migrations. Typically geographers are not satisfied with national averages or aggregates, since they often conceal sharp regional contrasts, and instead attempt to measure and depict these regional and local variations. In desert areas, for example, most of the population is concentrated in a few oases, which occupy a tiny fraction of the area. Minority ethnic, linguistic, or religious groups often show marked clustering, sometimes being concentrated in relatively homogeneous subregions. Some regions increase in population while others decline, these shifts often being accompanied by substantial migration flows. Some geographic studies are concerned mainly with spatial distribution, spatial mobility, or spatial diversity in relationship to environments or resources, all of which are often depicted on maps. Other studies are more concerned with fertility, mortality, population growth, and forecasting through the use of demographic models. Still others address questions of population policy. Interest in population geography in the second half of the 20th century has been heightened by the sharp and growing contrast between the economically developed countries—which have long experienced a demographic transition from high to low birth and death rates and thus to low rates of population increase—and the less-developed countries—in which death rates have declined dramatically but birth rates have remained high, resulting in rapid population growth that has posed extremely difficult problems.

Rural–urban shift in emphasis

The emphasis in human-geography studies has shifted to reflect how much less people are bound to the land as a result of the transformation in agriculture, increased industrialization, improved transportation and trade, changes in sources and forms of energy, urbanization, and the expansion of service industries. Location, accessibility, transportation and communication, life-styles, and economic and cultural factors increasingly have been recognized as key factors in the distribution of population. The diffusion of research in population geography from its early centres in Europe and North America to the world as a whole has emphasized greater culture-specific orientation in population studies and the contrasts in national policies between those advocating population limitation and others promoting rapid population increase. In the late 20th century large-scale studies of continental dimensions increasingly have given way to studies of small areas and even of individuals through household surveys, as the immense complexity and diversity of the processes of population growth and distribution have been recognized.

The field of population geography has developed mainly in the second half of the 20th century. Since 1956 the International Geographical Union has supported a special commission devoted to this field. Symposia have contributed greatly to international participation and progress. One of the reports in *Geography and Population* (1984), edited by John I. Clarke, noted the need to look at specific population problems of individual countries. Examples cited include the problems of decentralizing and de-urbanizing the population and of a growing proportion of the population that is very old (aged 75 and over) in some economically developed countries, the massive redistribution of population associated with settlement regroupings in Mozambique and Tanzania, the large influx of foreign migrant workers into the oil-producing countries of the Middle East, high population pressure and growth in the precarious deltaic environment of Bangladesh, high-mountain population pressure and out-migration in Nepal, the huge concentration of population in the capital of Mexico, the impact on population redistribution of development projects in The Sudan and of political polarization in Korea, the effects of pluralism and of a rapidly changing demographic pattern in many islands in the Pacific and Indian oceans and the Caribbean, the effectiveness of transmigration policies in Indonesia, and the out-migration of the labour force in Poland.

Economic geography. Basic to understanding the distribution of population is an awareness of how people make a living. The modern field of economic geography evolved out of the older commercial geography, which flourished in western Europe and North America in the late 19th and early 20th centuries with the growing interest in international trade. Economic geography, however, is far more complex than the earlier field in analyzing the characteristics of, differences among, and movements between areas in the production, exchange, and consumption of goods and services. Of special geographic interest are the localizations of economic activity as they have evolved historically within specific cultural and technological contexts, based on particular combinations of physical, biologic, and human resources, economic and political conditions, and interregional ties and movements. In studying how centres of a nation's iron and steel industry arose, for example, consideration must be given not only to the location and availability of raw materials—such as iron ore, coking coals, other fuels, limestone for flux, and scrap—but also such factors as the availability, qualifications, and cost of labour; distances and costs of reaching markets; effect of inertia in the form of past capital investments in plant and in social overhead (*e.g.,* transportation and communication systems); plant obsolescence; changing competitive positions through the introduction of new technologies; and even changes in foreign-exchange rates among the currencies of competing countries.

Classification of economic activities

Economic activities traditionally are classified as being either primary, secondary, or tertiary in nature. Primary production involves those activities that make direct use of natural resources and that are located close to such resources. They include agriculture, livestock raising, forestry, fishing and hunting, and mining and quarrying.

Secondary production involves processing primary products by manufacturing and related activities. Tertiary production refers to services—such as banking, marketing, and accounting—rather than to the production of goods. In advanced economies tertiary activities form an increasing portion of employment, as evidenced by the increasing percentage of people employed in services and by the growing number of office buildings in all of the world's larger cities. In addition to general works in economic geography—often devoted to a particular world region, country, or even individual city—many studies examine the principal subfields: agricultural geography, manufacturing (or industrial) geography, transportation, and trade (from the international movement of goods to retail trade in a single urban neighbourhood).

Five themes of cultural geography

Cultural and social geography. Five major themes characterize cultural geography: culture, culture area, cultural landscape, cultural history, and cultural ecology. The cultural geographer studies the distribution in space and time of cultures and the elements of culture, such as artifacts and tools, techniques, attitudes, customs, languages, and religious beliefs; cultural complexes in their spatial organization; the cultural landscape—*i.e.*, the association of human, biologic, and physical features on the surface of the Earth (especially as perceived visually), ranging from the natural landscape unaffected by humankind to the landscape as thoroughly transformed by human action; the evolution and succession of cultures and cultural elements, including the history of cultural origins and their areal diffusion; and the complex interrelationships and areal associations of culture and nature. The American geographer Carl O. Sauer was particularly creative in working the concepts and teaching of anthropology, archaeology, and sociology into geography.

Whereas the focus of cultural geography is more on traditional societies (though it is not restricted to them), social geography is more oriented toward urban problems in countries with advanced economies. Social geographers have been concerned particularly with the spatial aspects of disadvantaged groups (such as minorities, women, the aged, and the poor), of such social pathologies as crime and mental illness, and of inequality, social welfare, and housing.

Urban geography. Urban geography generally takes a much broader scope than cultural geography; it is a major field in the countries with well-developed economies and high levels of urbanization—western Europe, North America, Australia, and Japan. Among topics investigated are factors affecting the location of individual cities, urban systems as networks of settlement points, regional differences in urbanization, cities in relation to their tributary areas or spheres of influence, the hierarchy of central places, the characteristics of city-size regularities (such as the rule of ranking cities according to size), functional types of cities (economic classification), expansion of metropolitan areas, internal spatial structure of land use, urban transportation and areal patterns of commuting, social problems in cities, housing, and cities as growth poles. One of the seminal contributions to the discipline was made by the German geographer Walter Christaller, who in the first half of the 20th century studied the urban centres of southern Germany and demonstrated that there existed a hierarchy among these centres as well as spacing and size regularities in tributary areas.

Increased importance in the 20th century

Cities are often the locus of studies for many branches of geography—economic, social, or political. The relative role of urban geography has increased dramatically during the 20th century, as the proportion of the population living in urban areas has risen and as interest in the field of geography has shifted from the concern with agricultural geography (and its emphasis on physical factors, such as rainfall, temperature, and soil) to a focus on the burgeoning urban agglomerations as centres of economic, political, and social development and problems. Symbolic of this shift have been studies undertaken of the function of cities as office centres, in which offices are concentrated in skyscrapers that are packed together in central business districts for maximum accessibility to financial, legal, and marketing services both inside and outside the city. Although the development of high-speed express highways and of regional airports has tended to diffuse activities within metropolitan areas from central cities to suburban fringes—housing, manufacturing, and retail trade in particular have exhibited this tendency—the growth of business service activities and of office functions has tended to maintain high daytime populations in central areas, with workers often commuting from suburban communities and metropolitan fringes.

Political geography. Studies in political geography at the international level have been concerned with the organization of the world into states; with their larger political groupings into regional alliances on the one scale and their subordinate division into political-administrative units on another; with the functions, delimitation, and demarcation of boundaries; with the selection of capital sites; with the relation of core areas and peripheral areas; with metropolitan powers and colonies; with the bases of political power in terms of population, production, organization, and policy; with the relations among states, including international trade and aid; with international organizations; and with territorial waters, maritime boundaries, and the law of the sea. At the national scale, studies have been concerned with regionalism, including separatist movements and their bases, and increasingly with the areal analysis of voting patterns as they reflect regional interests. At the metropolitan level studies have been devoted to the political fragmentation of metropolitan areas into hundreds and even thousands of separate political bodies and to the rise of new organizational forms of metropolitan-wide bodies with taxing powers, such as water districts, transportation authorities, and voluntary planning associations.

Three areas of investigation

Medical geography. Three quite different types of investigations are included under the broad umbrella of medical geography. First, there are studies of the diffusion of infectious diseases from their centres of occurrence by specific pathways through space and time. These studies include the mapping of distribution of a certain disease (such as sleeping sickness in Africa) in relation to vectors in the transmission of the disease (such as the tsetse fly) or a series of geographic studies on the areal diffusion of infectious diseases, such as successive waves of a disease following particular outbreaks. An early classic study of the second type was made by the English physician John Snow, who plotted on a map the cases of cholera from an outbreak in 1854 in London and found that all had used water from a centrally placed public well that he determined was being contaminated by nearby latrines. Second are studies of the ecology of malnutrition in relation to medical problems. These include analyses of dietary deficiencies as they are related to general poverty or to specific inadequacies of the food supply, such as of proteins or vitamins. Third are surveys of health-care facilities—their optimum spatial location and allocation compared to their actual distribution—as, for example, the lack of facilities in certain districts of a city or in regions of a country or for disadvantaged segments of the population.

As it came to be recognized that the construction of man-made lakes was also creating unanticipated health hazards, researchers in medical geography have been called upon to determine how these problems arose and how they can be prevented in the future. The impoundment of Lake Volta in Ghana, for example, caused a precipitous rise in the number of cases of schistosomiasis among those who settled on the lakeshore; and this rise was then linked to the increase of the disease-carrying snail population along the shoreline. Atlases have come to provide useful summaries of the distribution of diseases and other causes of mortality.

Historical geography. Given the rich history of Europe, the successive transformations of the landscape, and changing geographic relationships through time, it is not surprising that interest in historical geography has a long record. The stages in the transformation of the natural landscape to the cultural landscape have been studied in many of the countries of western, southern, and eastern Europe. The famous Domesday Book—the survey of England taken in 1086 on the orders of William I the Conqueror to inventory the lands and resources of the

The Domesday Book

country—has become the most thoroughly investigated early geographic record. The British geographer H. Clifford Darby's monumental seven-volume geographic analysis of this inventory, published between 1952 and 1977, was a landmark in research in historical geography.

Historical geography is concerned with the geography of past periods or dates or with changes over time as the landscape evolves. The first aspect is a horizontal consideration of the areal patterns during specific periods in time; the other entails a vertical analysis of the process of change through time. The field of historical geography expanded significantly in the latter half of the 20th century as graduate training of a considerable number of specialists emerged, particularly those trained by Darby at the University of Cambridge and by the Canadian geographer Andrew H. Clark at the University of Wisconsin, Madison, as well as by their students and successors.

Regional geography. In contrast to the systematic fields of geography, which view particular categories of physical, biologic, or human phenomena as distributed over the globe, regional geography is concerned with the associations within regions of all or some of these elements, particularly in regard to their interrelationships and as they have evolved historically. The particular elements that give distinctive character to regions differ from place to place. High altitudes and steep slopes are the distinctive features of the Himalayas or Andes; forests are the features of the Amazon Basin, Siberia, and northern Canada; dryness and scanty vegetation, of the Sahara, the interior of Australia, and Central Asia; highly intensive agriculture, of the rice-producing areas of monsoonal Asia; commercial agriculture and livestock production, of the American Midwest; urbanized landscapes dotted with industries and connected by transport lines, of the belt from southern England through the Low Countries to the Ruhr district in Germany, of the manufacturing belt of the northeastern United States and adjacent Canada, and of the Pacific coast of Japan; ice sheets, of Greenland and Antarctica; and extensive grasslands used for grazing, of the dry margins of agriculture in the western United States, the interior of Argentina, southeastern Australia, and the steppes of Ukraine, Russia, and Central Asia. The distinguishing element may be surface configuration, ice and snow, climate, vegetation, or type of human activity—*i.e.,* pastoral, agricultural, industrial, or commercial.

Rise of the modern discipline

Although interest in how countries differ dates to the geographic studies of antiquity, systematic modern study of regional characteristics and differences dates from the late 19th and early 20th centuries and the pioneering work of Hettner in Germany, who emphasized the physical basis of regions, Vidal de La Blache in France, who appreciated especially the historical element in the evolution of a sense of region by its inhabitants, and Herbertson in England, who suggested the concept of natural regions. Throughout much of the 20th century regional geography has been extensively pursued, though at times it clearly has been subordinated to research in systematic branches of geography. The attempt to recognize and map ever smaller units of land that were homogeneous in some respect has led to minute regional subdivisions, particularly by scholars studying intensively their own countries or regions. The American Robert S. Platt was the first to make a sharp distinction between homogeneous regions, in which significant elements, either physical or human, were relatively uniform (such as the wheat belts of Canada, Argentina, and Australia), and regions of organization, where the unity derives from dynamic organizational forces from a centre (such as a political unit or the tributary area of a city).

One major question is how the world should be divided for regional treatment. For a long time the continents were considered to be the appropriate units. Regions of cultural similarity have gained increasing recognition over the years, the major divisions including Latin America, the Middle East and North Africa, the Mediterranean, Scandinavia, monsoonal Asia, and Central Asia. Vegetation and climatic belts display a close relationship with agricultural types and human activities and thus are also useful for many purposes.

BIBLIOGRAPHY

History of geography: PRESTON E. JAMES and GEOFFREY J. MARTIN, *All Possible Worlds: A History of Geographical Ideas,* 2nd ed. (1981), is a general history of the rise of the discipline of geography, with emphasis on the modern period. ROBERT E. DICKINSON and O.J.R. HOWARTH, *The Making of Geography* (1933, reprinted 1976), though dated in some respects, is still useful. GEORGE KISH (ed.), *A Source Book in Geography* (1978), is a handy annotated anthology. RICHARD HARTSHORNE, *The Nature of Geography* (1939, reprinted 1976), and *Perspective on the Nature of Geography* (1959, reissued 1968), provide a detailed examination of writings on the subject. J. OLIVER THOMSON, *History of Ancient Geography* (1948, reissued 1965), is a well-documented review of knowledge and theories. J.N.L. BAKER, *A History of Geographical Discovery and Exploration,* rev. ed. (1936, reissued 1967), remains the best single-volume summary of geographic explorations. See also the valuable bibliographies in *Geographers: Biobibliographical Studies* (irregular). For the modern period, see, for French and German work up to the 1960s, ROBERT E. DICKINSON, *The Makers of Modern Geography* (1969); on British geography, T.W. FREEMAN, *A History of Modern British Geography* (1980); on French geography, PAUL CLAVAL, *La Pensée géographique* (1972); on Soviet geography, I.P. GERASIMOV (ed.), *Soviet Geography,* trans. from Russian (1962); and on geography in the United States, PRESTON E. JAMES and CLARENCE F. JONES (eds.), *American Geography* (1954). Developments in 11 countries or regions in the latter part of the 20th century are discussed in R.J. JOHNSTON and PAUL CLAVAL (eds.), *Geography Since the Second World War* (1984).

Methodology: A standard introduction to cartography is ARTHUR H. ROBINSON *et al., Elements of Cartography,* 5th ed. (1984). Various aspects of cartography are covered in J.B. HARLEY and DAVID WOODWARD (eds.), *The History of Cartography* (1987–), an outstanding work, with one of six planned volumes published. More specialized are PHILLIP C. MUEHRCKE and JULIANA O. MUEHRKE, *Map Use: Reading, Analysis, and Interpretation,* 2nd ed. (1986); and MARK S. MONMONIER, *Computer-Assisted Cartography: Principles and Prospects* (1982). *The Times Atlas of the World,* 7th ed. (1985); RAND MCNALLY, *The New International Atlas* (1980, reissued 1987); and *National Geographic Atlas of the World,* 5th ed. (1981), are all examples of excellent cartography. Aerial photography is treated in C.P. LO, *Geographical Applications of Aerial Photography* (1976). Introductions to remote sensing are provided by BENJAMIN F. RICHASON, JR. (ed.), *Introduction to Remote Sensing of the Environment,* 2nd ed. (1983); and ROBERT K. HOLZ (ed.), *The Surveillant Science: Remote Sensing of the Environment,* 2nd ed. (1985). ROBERT N. COLWELL (ed.), *Manual of Remote Sensing,* 2nd ed., 2 vol. (1983), is advanced and comprehensive. Pioneering and influential volumes on quantitative methods include TORSTEN HÄGERSTRAND, *Innovation Diffusion as a Spatial Process* (1967; originally published in Swedish, 1953), a classic study; RICHARD J. CHORLEY and PETER HAGGETT (eds.), *Models in Geography* (1967); and PETER HAGGETT, ANDREW D. CLIFF, and ALLAN FREY, *Locational Analysis in Human Geography,* 2nd ed. (1977). See also JOHN F. LOUNSBURY and FRANK T. ALDRICH, *Introduction to Geographic Field Methods and Techniques,* 2nd ed. (1986).

Physical geography: (*General*): CUCHLAINE A.M. KING, *Physical Geography* (1980), provides coverage at local, regional, and continental scales. TOM L. MCKNIGHT, *Physical Geography: A Landscape Appreciation,* 2nd ed. (1987), is a succinct review. ARTHUR N. STRAHLER and ALAN H. STRAHLER, *Elements of Physical Geography,* 3rd ed. (1984), is comprehensive and well illustrated. MICHAEL J. CLARK, KENNETH J. GREGORY, and ANGELA M. GORNELL (eds.), *Horizons in Physical Geography* (1987), offers overviews of issues in the field.

(*Geomorphology*): KARL W. BUTZER, *Geomorphology from the Earth* (1976), is generally nontechnical. B.W. SPARKS, *Geomorphology,* 3rd ed. (1986), stresses deficiencies in theory. ARTHUR L. BLOOM, *Geomorphology: A Systematic Analysis of Late Cenozoic Landforms* (1978), is a thorough introduction. Works devoted to special processes include LUNA B. LEOPOLD, M. GORDON WOLMAN, and JOHN P. MILLER, *Fluvial Processes in Geomorphology* (1964); CLIFFORD EMBLETON and CUCHLAINE A.M. KING, *Glacial and Periglacial Geomorphology,* 2nd ed., 2 vol. (1975); CUCHLAINE A.M. KING, *Beaches and Coasts,* 2nd ed. (1972); and JULIUS BÜDEL, *Climatic Geomorphology* (1982; originally published in German, 1977).

(*Climatology*): Nontechnical introductions include HOWARD J. CRITCHFIELD, *General Climatology,* 4th ed. (1983); JOHN E. OLIVER and JOHN J. HIDORE, *Climatology* (1984); GLENN T. TREWARTHA and LYLE H. HORN, *An Introduction to Climate,* 5th ed. (1980); and ROGER G. BARRY and RICHARD J. CHORLEY, *Atmosphere, Weather, and Climate,* 5th ed. (1987). A comprehensive analysis is H.E. LANDSBERG (ed.), *World Survey of Climatology* (1969–), a multivolume series.

(*Biogeography*): Introductions are provided by JAMES H. BROWN and ARTHUR C. GIBSON, *Biogeography* (1983), with broad coverage and many references; PIERRE DANSEREAU, *Biogeography: An Ecological Perspective* (1957); I.G. SIMMONS, *Biogeography: Natural and Cultural* (1979), which focuses on human influence in the biosphere; and JOY TIVY, *Biogeography: A Study of Plants in the Ecosphere*, 2nd ed. (1982), which includes environmental and historical influences on plant distributions. Advanced specialized studies are contained in CARL TROLL (ed.), *Geo-Ecology of the Mountainous Regions of the Tropical Americas* (1968), and *Geoecology of the High-Mountain Regions of Eurasia* (1972).

(*Soil geography*): Brief introductions to the field include ROBERT M. BASILE, *A Geography of Soils* (1971); and DONALD STEILA, *The Geography of Soils: Formation, Distribution, and Management* (1976).

(*Resource management and environmental studies*): Good general introductions, with extensive bibliographies on environmental studies, include IAN R. MANNERS and MARVIN W. MIKESELL (eds.), *Perspectives on Environment* (1974); and KENNETH A. HAMMOND, GEORGE MACINKO, and WILMA B. FAIRCHILD (eds.), *Sourcebook on the Environment: A Guide to the Literature* (1978). The seminal contributions of Gilbert F. White to resource management and essays on related themes are contained in ROBERT W. KATES and IAN BURTON (eds.), *Geography, Resources, and Environment*, 2 vol. (1986). Other valuable works include ANDREW GOUDIE, *The Human Impact on the Natural Environment*, 2nd ed. (1986); GILBERT F. WHITE (ed.), *Natural Hazards* (1974); IAN BURTON, ROBERT W. KATES, and GILBERT F. WHITE, *The Environment as Hazard* (1978); ROBERT W. KATES, *Risk Assessment of Environmental Hazard* (1978); ANNE V. WHYTE and IAN BURTON (eds.), *Environmental Risk Assessment* (1980); THOMAS F. SAARINEN, DAVID SEAMON, and JAMES L. SELL (eds.), *Environmental Perception and Behavior* (1984); J.T. COPPOCK and B.S. DUFFIELD, *Recreation in the Countryside* (1975); and STEPHEN SMITH, *Recreation Geography* (1983).

Human geography: (*General*): General introductions to the field include PETER HAGGETT, *Geography*, rev. 3rd ed. (1983), emphasizing ecological and spatial approaches; JAN O.M. BROEK and JOHN W. WEBB, *A Geography of Mankind*, 3rd ed. (1978), a cultural and economic approach; and RHOADS MURPHEY, *Patterns on the Earth: An Introduction to Geography*, 4th ed. (1978), a historical and cultural approach within a regional framework.

(*Population geography*): A good international summary of approaches is JOHN I. CLARKE (ed.), *Geography and Population* (1984). Useful general reviews of the field include GEORGE A. SCHNELL and MARK S. MONMONIER, *The Study of Population* (1983), with world maps showing countries scaled by population size; and HUW R. JONES, *A Population Geography* (1981), with an extensive bibliography. Population pressures in relation to resources and resource depletion are treated in GARY L. PETERS and ROBERT P. LARKIN, *Population Geography*, 2nd ed. (1983); and WILBUR ZELINSKY, LESZEK A. KOSÍNSKI, and R. MANSELL PROTHERO (eds.), *Geography and a Crowding World* (1970), papers from a symposium on population pressures on physical and social resources in the developing lands. Population in relation to development is considered in PHILIP M. HAUSER (ed.), *World Population and Development* (1979). Migration is reviewed in LESZEK A. KOSÍNSKI and R. MANSELL PROTHERO (eds.), *People on the Move: Studies on Internal Migration* (1975).

(*Economic geography*): Brief introductions include HAROLD H. MCCARTY and JAMES B. LINDBERG, *A Preface to Economic Geography* (1966); and ROBERT B. MCNEE, *A Primer on Economic Geography* (1971). General treatments include RICHARD S. THOMAN and PETER B. CORBIN, *The Geography of Economic Activity*, 3rd ed. (1974); and JAMES O. WHEELER and PETER O. MULLER, *Economic Geography*, 2nd ed. (1986), with a North American orientation. Somewhat more advanced and theoretical are BRIAN J.L. BERRY, EDGAR C. CONKLING, and D. MICHAEL RAY, *The Geography of Economic Systems* (1976); PETER E. LLOYD and PETER DICKEN, *Location in Space: A Theoretical Approach to Economic Geography*, 2nd ed. (1977); and THOMAS J. WILBANKS, *Location and Well-Being: An Introduction to Economic Geography* (1980).

(*Cultural and social geography*): General treatments include PHILIP L. WAGNER and MARVIN W. MIKESELL (eds.), *Readings in Cultural Geography* (1962); J.E. SPENCER and WILLIAM L. THOMAS, *Introducing Cultural Geography*, 2nd ed. (1978); and TERRY G. JORDAN and LESTER ROWNTREE, *The Human Mosaic: A Thematic Introduction to Cultural Geography*, 4th ed. (1986). Two important collections of papers are WILLIAM L. THOMAS (ed.), *Man's Role in Changing the Face of the Earth* (1956, reissued in 2 vol., 1970); and JOHN LEIGHLY (ed.), *Land and Life: A Selection from the Writings of Carl Ortwin Sauer* (1963, reprinted 1974). General introductions to social geography include DAVID LEY, *A Social Geography of the City* (1983); and PAUL KNOX, *Urban Social Geography: An Introduction*, 2nd ed. (1987).

(*Urban geography*): Major works on worldwide urbanization include BRIAN J.L. BERRY, *Comparative Urbanization*, rev. and enl. 2nd ed. (1981); L.S. BOURNE, R. SINCLAIR, and K. DZIEWÓNSKI (eds.), *Urbanization and Settlement Systems* (1984); and STANLEY D. BRUNN and JACK F. WILLIAMS (eds.), *Cities of the World: World Regional Urban Development* (1983). North American cities are treated in TRUMAN ASA HARTSHORN, *Interpreting the City: An Urban Geography* (1980); STANLEY D. BRUNN and JAMES O. WHEELER (eds.), *The American Metropolitan System* (1980); RISA PALM, *The Geography of American Cities* (1981); MAURICE YEATES and BARRY GARNER, *The North American City*, 3rd ed. (1980); and PETER O. MULLER, *Contemporary Suburban America* (1981). JEAN GOTTMANN, *Megalopolis: The Urbanized Northeastern Seaboard of the United States* (1961, reissued 1969), is a classic study. A similar study for Canada is MAURICE YEATES, *Main Street: Windsor to Quebec City* (1975). British cities are more fully treated in HAROLD CARTER, *The Study of Urban Geography*, 3rd ed. (1981); and DAVID T. HERBERT and COLIN J. THOMAS, *Urban Geography* (1982). WALTER CHRISTALLER, *Central Places in Southern Germany* (1966; originally published in German, 1933), is a widely influential study.

(*Political geography*): ALAN D. BURNETT and PETER J. TAYLOR (eds.), *Political Studies from Spatial Perspectives: Anglo-American Essays on Political Geography* (1981), is an influential collection of papers; more diverse is PETER J. TAYLOR and JOHN HOUSE (eds.), *Political Geography* (1984). MARTIN IRA GLASSNER and HARM J. DE BLIJ, *Systematic Political Geography*, 3rd ed. (1980), is a widely used text. Special topics are treated by J.R.V. PRESCOTT, *Boundaries and Frontiers* (1978), and *The Political Geography of the Oceans* (1975); PETER J. TAYLOR and R.J. JOHNSTON, *The Geography of Elections* (1979); and JEAN GOTTMANN (ed.), *Centre and Periphery: Spatial Variation in Politics* (1980).

(*Medical geography*): Good introductions to medical geography include ANDREW T.A. LEARMONTH, *Patterns of Disease and Hunger* (1978); GERALD F. PYLE, *Applied Medical Geography* (1979); and the older but still valuable L. DUDLEY STAMP, *The Geography of Life and Death* (1964). Studies of infectious diseases are summarized in JACQUES M. MAY, *The Ecology of Human Disease* (1958); and JACQUES M. MAY (ed.), *Studies in Disease Ecology* (1961). For analyses of spatial aspects of health-care delivery systems, see JOHN EYLES and KEVIN J. WOODS, *The Social Geography of Medicine and Health* (1983), international in scope; DAVID R. PHILLIPS, *Contemporary Issues in the Geography of Health Care* (1981), a comparison of the British and American systems; and GARY W. SHANNON and G.E. ALAN DEVER, *Health Care Delivery: Spatial Perspectives* (1974), primarily American in focus. Diverse aspects are presented in the following conference papers or collected volumes: ANDREW T.A. LEARMONTH (ed.), *The Geography of Health* (1981); NEIL D. MCGLASHAN and JOHN R. BLUNDEN (eds.), *Geographical Aspects of Health* (1983); MELINDA S. MEADE (ed.), *Conceptual and Methodological Issues in Medical Geography* (1980); and GERALD F. PYLE (ed.), *New Directions in Medical Geography* (1979).

(*Historical geography*): An international overview of work in historical geography during the period 1945–70 is provided by ALAN R.H. BAKER (ed.), *Progress in Historical Geography* (1972). Research methods are discussed in ALAN R.H. BAKER and MARK BILLINGE (eds.), *Period and Place: Research Methods in Historical Geography* (1982); WILLIAM NORTON, *Historical Analysis in Geography* (1984); and ALAN R.H. BAKER and DEREK GREGORY (eds.), *Explorations in Historical Geography* (1984). H.C. DARBY (ed.), *A New Historical Geography of England* (1973); and R.A. DODGSHON and R.A. BUTLIN (eds.), *An Historical Geography of England and Wales* (1978), are valuable summaries. D.W. MEINIG, *The Shaping of America: A Geographical Perspective on 500 Years of History*, vol. 1, *Atlantic America, 1492–1800* (1986), is an important synthesis. JAMES R. GIBSON (ed.), *European Settlement and Development in North America* (1978); and ROBERT D. MITCHELL and PAUL A. GROVES (eds.), *North America: The Historical Geography of a Changing Continent* (1987), contain many valuable studies. See also RONALD E. GRIM, *Historical Geography of the United States: A Guide to Information Sources* (1982), which provides admirable coverage; and R. COLE HARRIS and JOHN WARKENTIN, *Canada Before Confederation: A Study in Historical Geography* (1974), which covers the early period in Canada. Prehistory is discussed in KARL W. BUTZER, *Environment and Archeology* (1971), and *Archaeology as Human Ecology* (1982).

Regional geography: Some useful treatments covering the large regions of the world include C. LANGDON WHITE, EDWIN J. FOSCUE, and TOM L. MCKNIGHT, *Regional Geography of Anglo-America*, 6th ed. (1985); J.H. PATERSON and CLARENCE W. OLMSTEAD, *North America: A Geography of Canada and the United*

States, 7th ed. (1984); J. WREFORD WATSON, *North America,* 2nd ed. (1968); *Studies in Canadian Geography,* 6 vol. (1972); PRESTON E. JAMES, C.W. MINKEL, and EILEEN W. JAMES, *Latin America,* 5th ed. (1986); ROBERT C. WEST and JOHN P. AUGELLI, *Middle America,* 2nd ed. (1976); GEORGE W. HOFFMAN (ed.), *A Geography of Europe,* 5th ed. (1983); PAUL E. LYDOLPH, *Geography of the U.S.S.R.: Topical Analysis* (1979); R.J. JOHNSTON and J.C. DORNKAMP (eds.), *The Changing Geography of the United Kingdom* (1982); WILLIAM A. HANCE, *The Geography of Modern Africa,* 2nd ed. (1975); R.J. HARRISON CHURCH *et al., Africa and the Islands,* 4th ed. (1977); A.T. GROVE, *Africa,* 3rd ed. (1978); W.B. FISHER, *The Middle East: A Physical, Social, and Regional Geography,* 7th ed. rev. (1978); NORTON GINSBURG, *The Pattern of Asia* (1958); J.E. SPENCER and WILLIAM L. THOMAS, *Asia, East by South: A Cultural Geography,* 2nd ed. (1971); O.H.K. SPATE and ANDREW T.A. LEARMONTH, *India and Pakistan: A General and Regional Geography,* 3rd ed. rev. (1967); and D.N. JEANS (ed.), *Australia: A Geography* (1977).

Bibliographic guides: Further bibliographic information may be found in STEPHEN GODDARD (ed.), *A Guide to Information Sources in the Geographical Sciences* (1983); J. GORDON BREWER, *The Literature of Geography: A Guide to Its Organisation and Use,* 2nd ed. (1978), textual discussion with some emphasis on British works; and CHAUNCY D. HARRIS, *Bibliography of Geography,* pt. 1, *Introduction to General Aids* (1976), and pt. 2, *Regional,* vol. 1, *The United States of America* (1984). CHAUNCY D. HARRIS *et al.* (eds.), *A Geographical Bibliography for American Libraries* (1985), focuses on the period 1970–84, though earlier works are also included; while GORDON R. LEWTHWAITE, EDWARD T. PRICE, JR., and HAROLD A. WINTERS (comps. and eds.), *A Geographical Bibliography for American College Libraries,* rev. ed. (1970), is particularly useful for the period before 1970. *Geographical Abstracts* (monthly), is the fullest bibliography in English and is especially strong in systematic fields of geography. *Current Geographical Publications* (monthly) concentrates on American publications. *Bibliographie Géographie Internationale* (quarterly) is international in scope and the best bibliography on regional geography.

Periodicals: Leading international journals in English include *Annals of the Association of American Geographers* (quarterly); *Canadian Geographer/Geographe Canadien* (quarterly); *Geographical Journal* (3/yr.); *Geographical Review* (quarterly); and *Transactions of the Institute of British Geographers* (quarterly). See also CHAUNCY D. HARRIS, *Annotated World List of Selected Current Geographical Serials,* 4th ed. (1980), which annotates 443 of the most valuable current serials from 72 countries.

Dictionaries: AUDREY N. CLARK, *Longman Dictionary of Geography: Human and Physical* (1985), offers good brief definitions of terms in all fields. R.J. JOHNSTON (ed.), *The Dictionary of Human Geography,* 2nd ed. (1986), includes advanced, signed definitions. ANDREW GOUDIE *et al., The Encyclopaedic Dictionary of Physical Geography* (1985), is authoritative, with a focus on concepts.

(C.D.H.)

Geometry

Geometry is the branch of mathematics concerned with the shape of individual objects, spatial relationships among various objects, and the properties of surrounding space. It is one of the oldest branches of mathematics. The earliest known unambiguous examples of written records—dating from Egypt and Mesopotamia about 3100 BC—demonstrate that ancient peoples had already begun to devise mathematical rules and techniques useful for surveying land areas, constructing buildings, and measuring storage containers. Beginning about the 6th century BC, the Greeks gathered and extended this practical knowledge and from it generalized the abstract subject now known as geometry, from the combination of the Greek words *geo* ("Earth") and *metron* ("measure") for the measurement of the Earth. Thales of Miletus (fl. *c.* 600 BC), one of the legendary Seven Wise Men from ancient Greece, is credited with fashioning the first geometric, or even mathematical, proofs—that is, employing logical reasoning to think about idealized, or abstract, entities in order to establish general, or theoretical, results—in contrast to earlier "recipes" that applied to specific examples from direct experience. Over the following centuries, various Greeks refined and developed the subject until, by the time of Plato and Aristotle in the 4th century BC, geometry had become the model for reasoning, influencing first the development of logic and eventually all of science.

While Greek mathematical knowledge all but disappeared from Europe during the Middle Ages, the Islāmic world preserved Greek results and made further mathematical progress, particularly in trigonometry and algebra. As ancient Greek and Islāmic mathematics filtered back into Europe at the end of the Middle Ages, it was realized that geometry need not be limited to the study of flat surfaces (plane geometry) and rigid three-dimensional objects (solid geometry) but that even the most abstract thoughts and images might be represented and developed in geometric terms. In particular, algebra was combined with geometry in 17th-century France to create analytic geometry. In addition to applying algebraic notation and techniques to common geometric figures, it revealed the hitherto unknown "shape" of mathematical expressions, including expressions that contain higher powers on the variable, or multivariable, terms. Such work led directly to the discovery of differential calculus and, from the mid-18th century to the early 19th century, to the development of differential geometry. Seventeenth-century France also saw the creation of projective geometry, although it was then rather neglected until the 19th century.

This article begins with an extensive historical treatment. In addition to describing some of the achievements of the ancient Greeks, this historical overview examines some applications of geometry to astronomy, cartography, and painting from classical Greece through the Islāmic world to Renaissance Europe and beyond. It concludes with a brief discussion of non-Euclidean geometries. The article then proceeds to descriptions of the origins, fundamental concepts, and distinctive procedures of the major branches of geometry. Additional information about the development and history of geometry can be found in MATHEMATICS, THE FOUNDATIONS OF; MATHEMATICS, THE HISTORY OF; ALGEBRA; ANALYSIS; and TRIGONOMETRY.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 10/21, 10/22, and 10/23, and the *Index.*

This article is divided into the following sections:

HISTORY OF GEOMETRY

Ancient geometry: practical and empirical

The origin of geometry lies in the concerns of everyday life. The traditional account, preserved in Herodotus's *History* (5th century BC), credits the Egyptians with inventing surveying in order to reestablish property values after the annual flood of the Nile. Similarly, eagerness to know the volumes of solid figures derived from the need to evaluate tribute, store oil and grain, and build dams and pyramids. Even the three abstruse geometrical problems of ancient times—to double a cube, trisect an angle, and square a circle, all of which are discussed below—probably arose from practical matters, from religious ritual, timekeeping, and construction, respectively, in pre-Greek societies of the Mediterranean. And the main subject of later Greek geometry, the theory of conic sections, owed its general importance, and perhaps also its origin, to its application to optics and astronomy.

While many ancient individuals, known and unknown, contributed to the subject, none equaled the impact of Euclid and his *Elements* of geometry, a book now 2,300 years old and the object of as much painful and painstaking study as the Bible. Much less is known about Euclid, however, than about Moses. In fact, the only thing known with a fair degree of confidence is that Euclid taught at the Library of Alexandria during the reign of Ptolemy I (323–285/83 BC). Euclid wrote not only on geometry but also on astronomy and optics and perhaps also on mechanics and music. Only the *Elements,* which was extensively copied and translated, has survived intact.

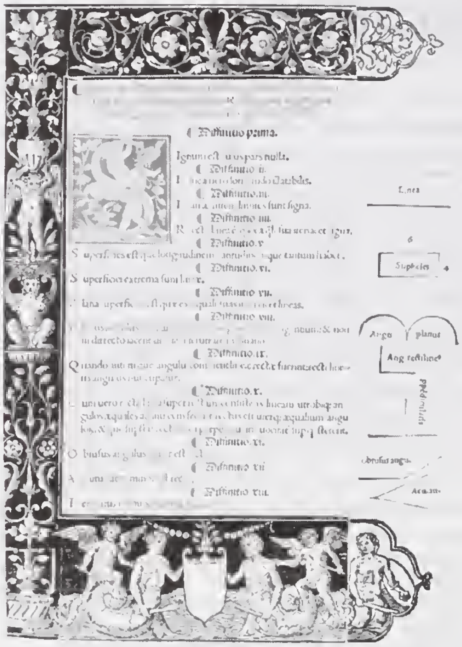


Figure 2: A page from a printed edition of Euclid's *Elements.*
Art Resource New York

Euclid's *Elements* was so complete and clearly written that it literally obliterated the work of his predecessors. What is known about Greek geometry before him comes primarily from bits quoted by Plato and Aristotle and by later mathematicians and commentators. Among other precious items they preserved are some results and the general approach of Pythagoras (*c.* 580–*c.* 500 BC) and his followers. The Pythagoreans convinced themselves that all things are, or owe their relationships to, numbers. The doc-

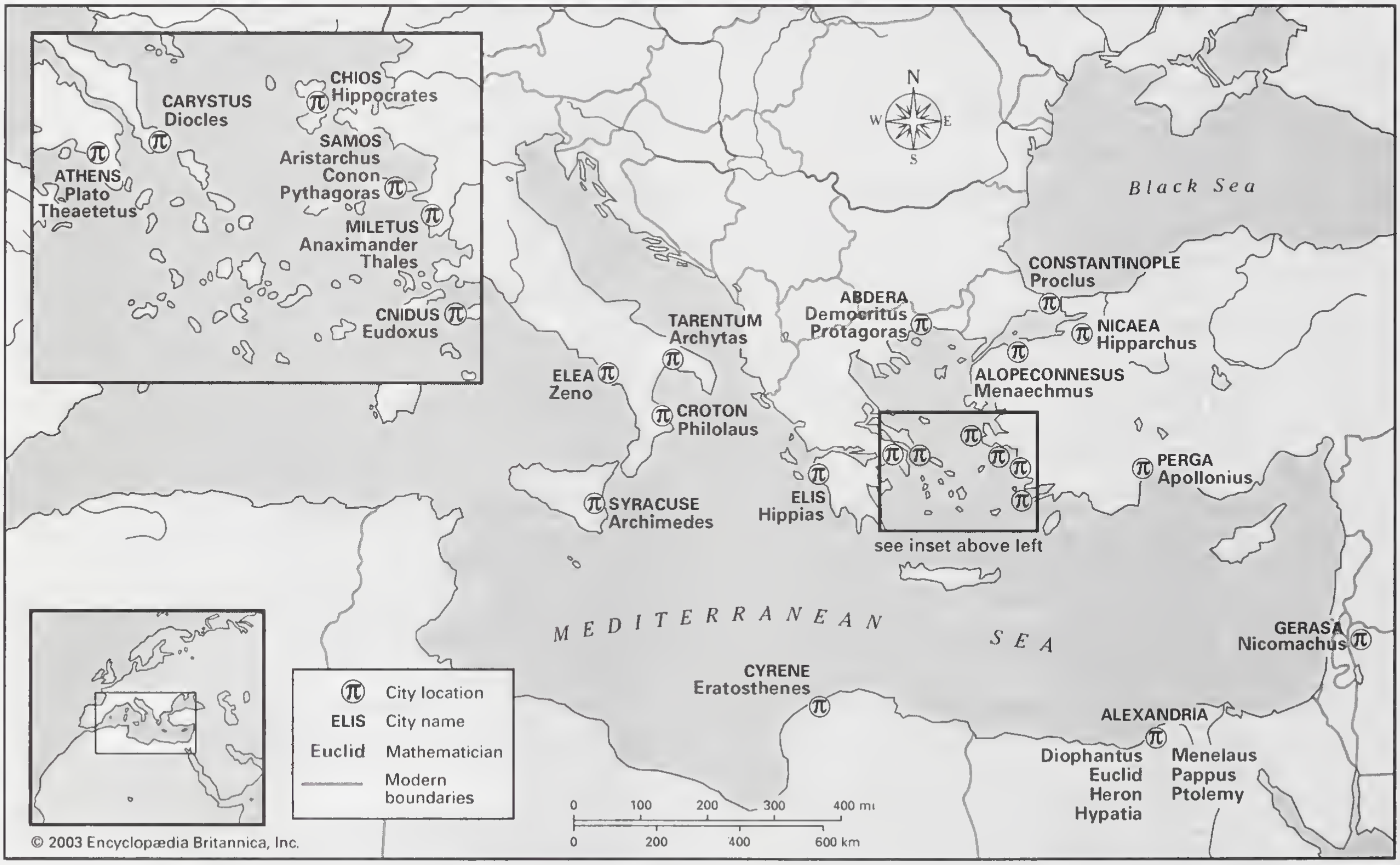


Figure 1: *Mathematicians of the Greco-Roman world.*
This map spans a millennium of prominent Greco-Roman mathematicians, from Thales of Miletus (*c.* 600 BC) to Hypatia of Alexandria (*c.* AD 400).

trine gave mathematics supreme importance in the investigation and understanding of the world. Plato developed a similar view, and philosophers influenced by Pythagoras or Plato often wrote ecstatically about geometry as the key to the interpretation of the universe. Thus ancient geometry gained an association with the sublime to complement its earthy origins and its reputation as the exemplar of precise reasoning.

FINDING THE RIGHT ANGLE

Egyptian rope pullers

Ancient builders and surveyors needed to be able to construct right angles in the field on demand. The method employed by the Egyptians earned them the name "rope pullers" in Greece, apparently because they employed a rope for laying out their construction guidelines. One way that they could have employed a rope to construct right triangles was to mark a looped rope with knots so that, when held at the knots and pulled tight, the rope must form a right triangle. The simplest way to perform the trick is to take a rope that is 12 units long, make a knot 3 units from one end and another 5 units from the other end, and then knot the ends together to form a loop, as shown in Figure 3. However, the Egyptian scribes have not left us instructions about these procedures, much less any hint that they knew how to generalize them to obtain the Pythagorean theorem: the square on the line opposite the right angle equals the sum of the squares on the other two sides. Similarly, the Vedic scriptures of ancient India contain sections called *sulvasutras,* or "rules of the rope," for the exact positioning of sacrificial altars. The required right angles were made by ropes marked to give the triads (3, 4, 5) and (5, 12, 13).

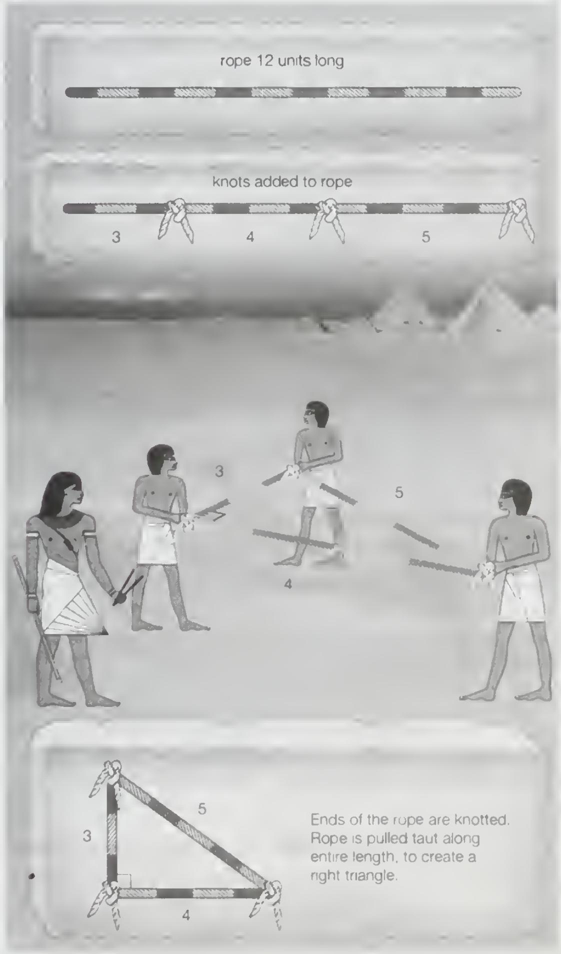

rope 12 units long
knots added to rope
3
4
5
3
5
4
5
3
4
Ends of the rope are knotted. Rope is pulled taut along entire length, to create a right triangle.

Figure 3. Egyptian rope pullers constructing a right triangle.

In Babylonian clay tablets (*c.* 1700–1500 BC) modern historians have discovered problems whose solutions indicate that the Pythagorean theorem and some special triads were known more than a thousand years before Euclid. A right triangle made at random, however, is very unlikely to have all its sides measurable by the same unit—that is, every side a whole-number multiple of some common unit of measurement. This fact, which came as a shock when discovered by the Pythagoreans, gave rise to the concept and theory of incommensurability.

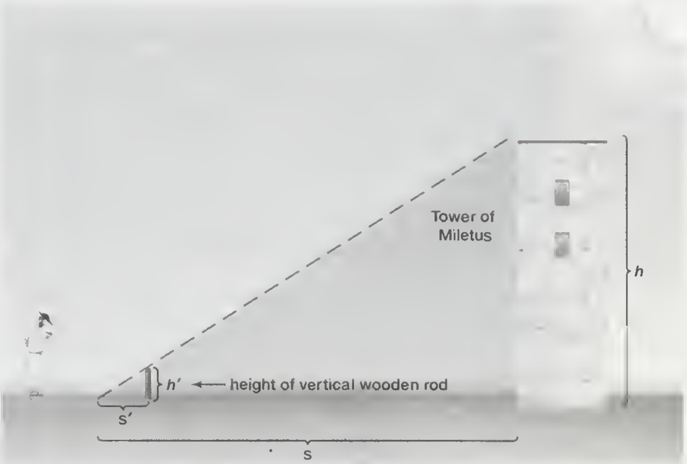


Figure 4: *Caricature of Thales measuring the height of a tower in Miletus.*
By aligning the shadows cast by the tower and the rod, the large triangle with sides *s* and *h* is similar to the small triangle with corresponding sides *s′* and *h′*. Hence, the ratio of *h* to *s* is equal to the ratio of *h′* to *s′*. The lengths of *s*, *s′*, and *h′* can be measured and used to calculate *h*, the height of the tower.

LOCATING THE INACCESSIBLE

By ancient tradition, Thales of Miletus, who lived before Pythagoras in the 6th century BC, invented a way to measure inaccessible heights (see Figure 4), such as the Egyptian pyramids. Although none of his writings survives, Thales may well have known about a Babylonian observation that for similar triangles (triangles having the same shape but not necessarily the same size) the length of each corresponding side is increased (or decreased) by the same multiple. The ancient Chinese arrived at measures of inaccessible heights and distances by another route, using "complementary" rectangles, as seen in Figure 5, which can be shown to give results equivalent to those of the Greek method involving triangles.

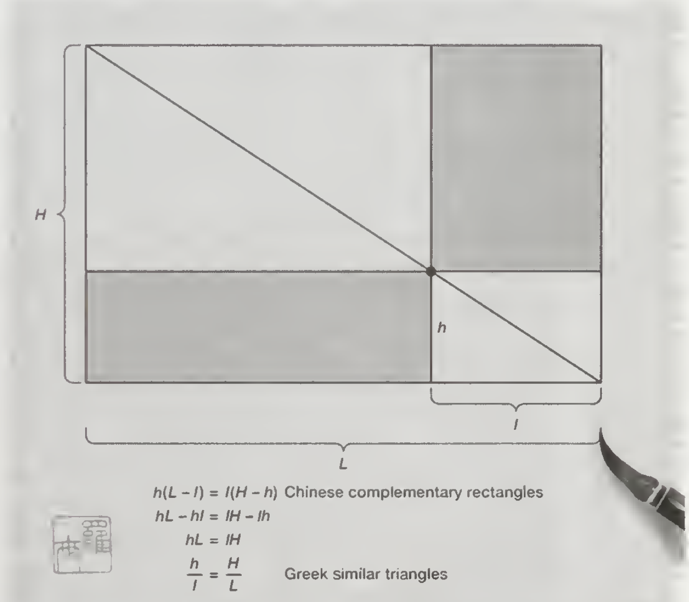

H
h
l
L
$h(L - l) = l(H - h)$ Chinese complementary rectangles
$hL - hl = lH - lh$
$hL = lH$
$\frac{h}{l} = \frac{H}{L}$ Greek similar triangles

Figure 5: An illustration of the equivalence of the Chinese complementary rectangles theorem and the Greek similar triangles theorem.

ESTIMATING THE WEALTH

Early estimates of π

A Babylonian cuneiform tablet written some 3,500 years ago treats problems about dams, wells, water clocks, and excavations. It also has an exercise on circular enclosures with an implied value of $\pi = 3$. The contractor for King Solomon's swimming pool, who made a pond 10 cubits across and 30 cubits around (1 Kings 7:23), used the same value. However, the Hebrews should have taken their π from the Egyptians before crossing the Red Sea, for the Rhind papyrus (*c.* 2000 BC; our principal source for ancient Egyptian mathematics) implies a more accurate value: $\pi = 3.1605$.

Knowledge of the area of a circle was of practical value to the officials who kept track of the pharaoh's tribute as well as to the builders of altars and swimming pools. Ahmes, the scribe who copied and annotated the Rhind papyrus (*c.* 1650 BC), has much to say about cylindrical granaries and pyramids, whole and truncated. He could calculate their volumes, and, as appears from his taking the Egyptian *seked,* the horizontal distance associated with a vertical rise of one cubit, as the defining quantity for the pyramid's slope, he knew something about similar triangles.

Ancient geometry: abstract and applied

THE THREE CLASSICAL PROBLEMS

Euclidean construction tools

In addition to proving mathematical theorems, ancient mathematicians constructed various geometrical objects. Euclid arbitrarily restricted the tools of construction to a straightedge (an unmarked ruler) and a compass. The restriction made three problems of particular interest (to double a cube, to trisect an arbitrary angle, and to square a circle) very difficult—in fact, impossible. Various methods of construction using other means were devised in the classical period, and efforts, always unsuccessful, using straightedge and compass persisted for the next 2,000 years. In 1837 the French mathematician Pierre Laurent Wantzel proved that doubling the cube and trisecting the angle are impossible, and in 1880 the German mathematician Ferdinand von Lindemann showed that squaring the circle is impossible, as a consequence of his proof that π is a transcendental number.

Doubling the cube. The Vedic scriptures made the cube the most advisable form of altar for anyone who wanted to supplicate in the same place twice. The rules of ritual required that the altar for the second plea have the same shape but twice the volume of the first. If the sides of the original and derived altars are a and b, respectively, then $b^3 = 2a^3$. The problem came to the Greeks together with its ceremonial content. An oracle disclosed that the citizens of Delos could free themselves of a plague merely by replacing an existing altar by one twice its size. The Delians applied to Plato. He replied that the oracle did not mean that the gods wanted a larger altar but that they had intended "to shame the Greeks for their neglect of mathematics and their contempt for geometry." With this blend of Vedic practice, Greek myth, and academic manipulation, the problem of the duplication of the cube took a leading place in the formation of Greek geometry.

Hippocrates of Chios, who wrote an early *Elements* about 450 BC, took the first steps in cracking the altar problem. He reduced the duplication to finding two mean proportionals between 1 and 2, that is, to finding lines x and y in the ratio $1{:}x = x{:}y = y{:}2$. After the intervention of the Delian oracle, several geometers around Plato's Academy found complicated ways of generating mean proportionals.

A few generations later, Eratosthenes of Cyrene (*c.* 276–*c.* 194 BC) devised a simple instrument with moving parts that could produce approximate mean proportionals.

Trisecting the angle. The Egyptians told time at night by the rising of 12 *asterisms* (constellations), each requiring on average two hours to rise. In order to obtain more convenient intervals, the Egyptians subdivided each of their *asterisms* into three parts, or *decans*. That presented the problem of trisection. It is not known whether the second celebrated problem of archaic Greek geometry, the trisection of any given angle, arose from the difficulty of the *decan,* but it is likely that it came from some problem in angular measure.

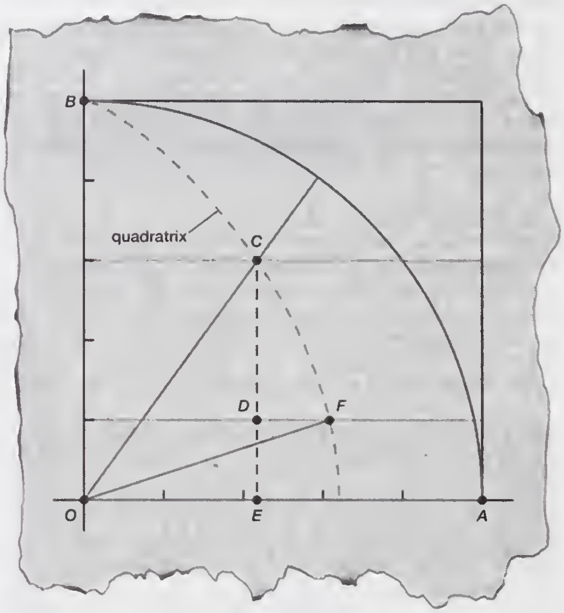


Figure 6: Quadratrix of Hippias.

Quadratrix. Several geometers of Plato's time tried their hands at trisection. Although no one succeeded in finding a solution with straightedge and compass, they did succeed with a mechanical device and by a trick. The mechanical device, perhaps never built, creates what the ancient geometers called a quadratrix. Invented by a geometer known as Hippias of Elis (fl. 5th century BC), the quadratrix is a curve traced by the point of intersection between two moving lines, one rotating uniformly through a right angle, the other gliding uniformly parallel to itself, as shown in Figure 6. Starting from a horizontal position, a line segment (OA) is rotated at a constant rate through a right angle around one of its endpoints (O) at the same time the segment (OA) also glides uniformly through a vertical distance equal to the segment's length. Because both the angle rotation and the vertical displacement are produced by uniform motion, each moves through the same fraction of its entire journey in the same time. Hence, finding some proportion (say one-third) for a given angle (here $\angle COA$) is simple: find the equal proportion for vertical displacement of the point on the quadratrix at which the two segments intersect (C), locate the point (F) on the quadratrix at that height (one-third of the original height in this example), and then draw the new angle ($\angle FOA$) through that point.

Neusis. The trick for trisection is an application of what the Greeks called *neusis*, a maneuvering of a measured length into a special position to complete a geometrical figure. A late version of its use, ascribed to Archimedes (*c.* 285–212/211 BC), exemplifies the method of angle trisection. (See Figure 7.)

1. Given $\angle AOB$, draw the circle with centre at O through the points A and B. Thus, OA and OB are radii of the circle and $OA = OB$.

2. Extend the ray AO indefinitely.

3. Now take a straightedge marked with the length of the circle's radius and maneuver it (this is the *neusis*) into position to draw a line segment from B through a point C on the circle to a point D on the ray AO such that CD is equal to the circle's radius; that is, $CD = OC = OB = OA$.

Bridge of Asses

4. By the "Bridge of Asses" (described below), $\angle CDO = \angle COD$ and $\angle OCB = \angle OBC$.

5. $\angle AOB = \angle ODC + \angle OBC$, because $\angle AOB$ is an angle external to ΔDOB and an external angle equals the sum of the opposite interior angles ($\angle AOB + \angle BOD = 180° = \angle BOD + \angle ODB + \angle DBO$).

6. $\angle OBC = \angle OCB$ (by step 4) $= \angle ODC + \angle COD$ (by step 5) $= 2\angle ODC$ (by step 4).

7. Substituting $2\angle ODC$ for $\angle OBC$ in step 5 and simplifying, $\angle AOB = 3\angle ODC$. Hence $\angle ODC$ is one-third the original angle, as required.

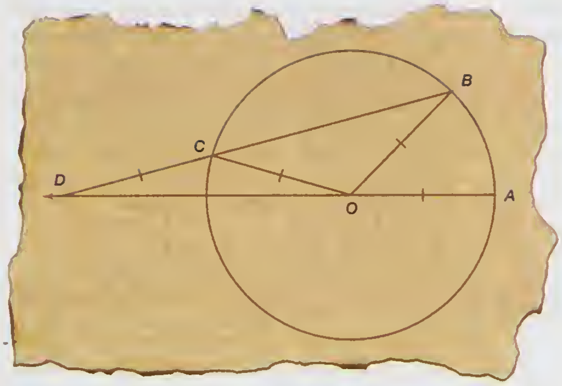


Figure 7: Archimedes' method of angle trisection.

Squaring the circle. The pre-Euclidean Greek geometers transformed the practical problem of determining the area of a circle into a tool of discovery. Two approaches can be distinguished: Hippocrates' dodge of substituting one problem for another; and the technique known in modern times as the "method of exhaustion" and attributed by its greatest practitioner, Archimedes, to Plato's student Eudoxus of Cnidus (*c.* 390–*c.* 340 BC).

Quadrature of lunes. While not able to square the circle, Hippocrates did demonstrate the quadratures of lunes; that is, he showed that the crescent moon-shaped area between two intersecting circular arcs could be expressed exactly as a rectilinear area, or quadrature. In the following simple case, shown in Figure 8, two lunes developed around the sides of a right triangle have a combined area equal to that of the triangle.

1. Starting with the right ΔABC, draw a circle whose diameter coincides with AB (side c), the hypotenuse. Because any right triangle drawn with a circle's diameter for its hypotenuse must be inscribed within the circle, C must be on the circle.

2. Draw semicircles with diameters AC (side b) and BC (side a).

3. Label the resulting lunes L_1 and L_2 and the resulting segments S_1 and S_2, as indicated in the figure.

4. Now the sum of the lunes (L_1 and L_2) must equal the sum of the semicircles ($L_1 + S_1$ and $L_2 + S_2$) containing them minus the two segments (S_1 and S_2). Thus, $L_1 + L_2 = (\pi/2)(b/2)^2 - S_1 + (\pi/2)(a/2)^2 - S_2$ (since the area of a circle is π times the square of the radius).

5. The sum of the segments (S_1 and S_2) equals the area of the semicircle based on AB minus the area of the triangle. Thus, $S_1 + S_2 = (\pi/2)(c/2)^2 - \Delta ABC$.

6. Substituting the expression in step 5 into step 4 and factoring out common terms, $L_1 + L_2 = (\pi/8)(a^2 + b^2 - c^2) + \Delta ABC$.

7. Since $\angle ACB = 90°$, $a^2 + b^2 - c^2 = 0$, by the Pythagorean theorem. Thus, $L_1 + L_2 = \Delta ABC$.

Hippocrates managed to square several sorts of lunes, some on arcs greater and less than semicircles, and he intimated, though he may not have believed, that his method could square an entire circle. At the end of the classical age, Boethius (*c.* AD 470–524), whose Latin translations of snippets of Euclid would keep the light of geometry flickering for half a millennium, mentioned that someone had accomplished the squaring of the circle. Whether the unknown genius used lunes or some other method is not known, since for lack of space Boethius did not give the demonstration. He thus transmitted the challenge of the quadrature of the circle together with fragments of geometry apparently useful in performing it. Europeans kept at the hapless task well into the Enlightenment. Finally, in 1775, the Paris Academy of Sciences, fed up with the task of spotting the fallacies in solutions submitted to it, refused to have anything further to do with circle squarers.

Circle
squares

Method of exhaustion. The method of exhaustion as developed by Eudoxus approximates a curve or surface by using polygons with calculable perimeters and areas. As the number of sides of a regular polygon inscribed in a circle increases indefinitely, its perimeter and area "exhaust," or take up, the circumference and area of the circle to within any assignable error of length or area, however small. In Archimedes' usage, the method of exhaustion produced upper and lower bounds for the value of π, the ratio of any circle's circumference to its diameter. This he accomplished by inscribing a polygon within a circle, and circumscribing a polygon around it as well, thereby bounding the circle's circumference between the polygons' calculable perimeters (see Figure 9). He used polygons with 96 sides and thus bound π between 310/71 and 31/7.

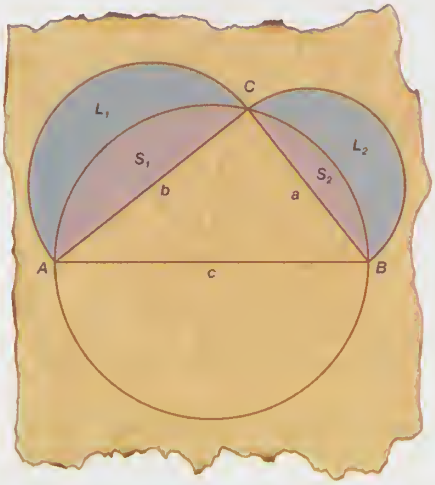


Figure 8: Quadrature of the lune.

IDEALIZATION AND PROOF

The last great Platonist and Euclidean commentator of antiquity, Proclus (*c.* AD 410–485), attributed to the inex-

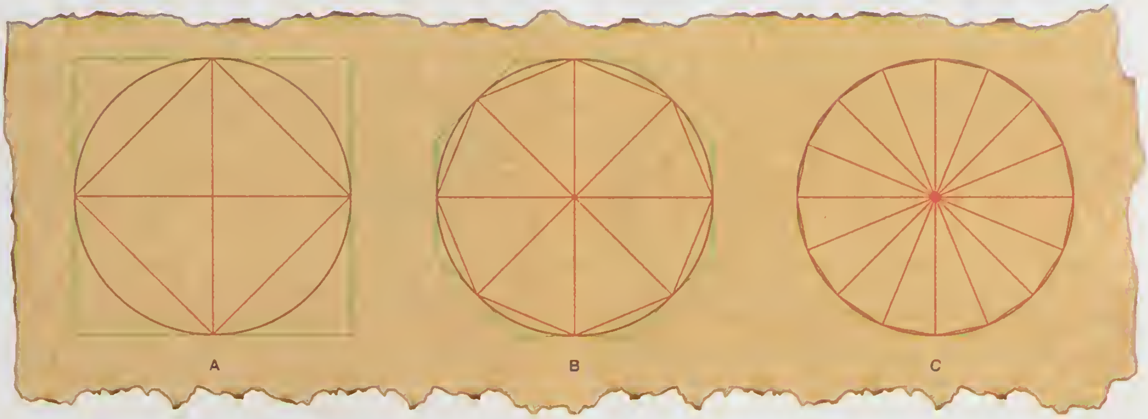


Figure 9: *Method of exhaustion.*
Archimedes' method of approximating the value of π is illustrated for polynomials with 4, 8, and 16 sides, respectively.

haustible Thales the discovery of the far-from-obvious proposition that even apparently obvious propositions need proof. Proclus referred especially to Thales' theorem that in an isosceles triangle the angles opposite the equal sides are equal—Euclid's fifth proposition in the first book of his *Elements*. In the Middle Ages the theorem was named the Bridge of Asses (Latin: Pons Asinorum), possibly for students who, clearly not destined to cross over into more abstract mathematics, had difficulty understanding the proof—or even the need for the proof. An alternative name was *Elefuga*, which Roger Bacon, writing circa AD 1250, derived from Greek words indicating "escape from misery." Medieval schoolboys did not usually go beyond the Bridge of Asses, which thus marked their last obstruction before liberation from the *Elements*.

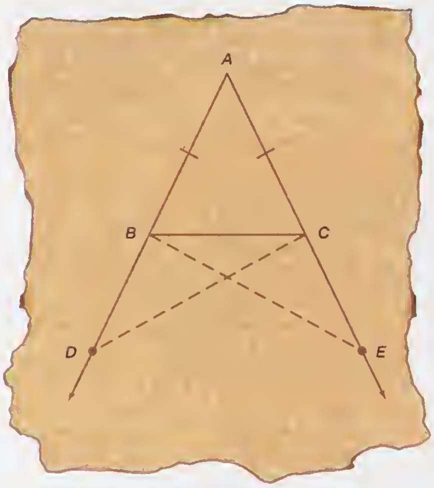


Figure 10: Bridge of Asses.

As shown in Figure 10, the Bridge of Asses may also have received its name from the figure described by Euclid for his proof:

1. We are given that ΔABC is an isosceles triangle—that is, that $AB = AC$.

2. Extend sides AB and AC indefinitely away from A.

3. With a compass centred on A and open to a distance larger than AB, mark off AD on AB extended and AE on AC extended so that $AD = AE$.

4. $\angle DAC = \angle EAB$, because it is the same angle.

5. Therefore, $\Delta DAC \cong \Delta EAB$; that is, all the corresponding sides and angles of the two triangles are equal. By imagining one triangle to be superimposed on another, Euclid argued that the two are congruent if two sides and the included angle of one triangle are equal to the corresponding two sides and included angle of the other triangle (known as the side-angle-side theorem).

6. Therefore, $\angle ADC = \angle AEB$ and $DC = EB$, by step 5.

7. Now $BD = CE$ because $BD = AD - AB$, $CE = AE - AC$, $AB = AC$, and $AD = AE$, all by construction.

8. $\Delta BDC \cong \Delta CEB$, by the side-angle-side theorem of step 5.

9. Therefore, $\angle DBC = \angle ECB$, by step 8.

10. Hence, $\angle ABC = \angle ACB$ because $\angle ABC = 180° - \angle DBC$ and $\angle ACB = 180° - \angle ECB$.

The ancient Greek geometers soon followed Thales over the Bridge of Asses. In the 5th century BC the philosopher-mathematician Democritus (*c.* 460–*c.* 370 BC) declared that his geometry excelled all the knowledge of the Egyptian rope pullers because he could prove what he claimed. By the time of Plato, geometers customarily proved their propositions. Their compulsion and the multiplication of theorems it produced fit perfectly with the endless questioning of Socrates and the uncompromising logic of Aristotle. Perhaps the origin, and certainly the exercise, of the peculiarly Greek method of mathematical proof should be sought in the same social setting that gave rise to the practice of philosophy—that is, the Greek *polis*. There citizens learned the skills of a governing class, and the wealthier among them enjoyed the leisure to engage their minds as they pleased, however useless the result, while slaves attended to the necessities of life. Greek society could support the transformation of geometry from a practical art to a deductive science. Despite its rigour, however, Greek geometry does not satisfy the demands of the modern systematist. Euclid himself sometimes appeals to inferences drawn from an intuitive grasp of concepts such as point and line or inside and outside, uses superposition, and so on. It took more than 2,000 years to purge the *Elements* of what pure deductivists deemed imperfections.

The first deductive science

THE EUCLIDEAN SYNTHESIS

Euclid, in keeping with the self-conscious logic of Aristotle, began the first of his 13 books of the *Elements* with sets of definitions ("a line is breadthless length"), common notions ("the whole is greater than the part"), and axioms, or postulates ("all right angles are equal"). Of this preliminary matter, the fifth and last postulate, which states a sufficient condition that two straight lines meet if sufficiently extended, has received by far the greatest attention. In effect it defines parallelism. Many later geometers tried to prove the fifth postulate using other parts of the *Elements*. Euclid saw farther, for coherent geometries (known as non-Euclidean geometries) can be produced by replacing the fifth postulate with other postulates that contradict Euclid's choice.

The first six books contain most of what Euclid delivers about plane geometry. Book I presents many propositions doubtless discovered by his predecessors, from Thales' theorem (Euclid's fifth proposition), discussed above, to the Pythagorean theorem, with which the book effectively ends.

Book VI applies the theory of proportion from Book V to similar figures and presents the geometrical solution to quadratic equations. As usual, some of it is older than Euclid. Books VII–X, which concern various sorts of numbers, especially primes, and various sorts of ratios, are seldom studied now, despite the importance of the masterful Book X, with its elaborate classification of incommensurable magnitudes, to the later development of Greek geometry.

Books XI–XIII deal with solids: XI contains theorems about the intersection of planes and of lines and planes and theorems about the volumes of parallelepipeds (solids with parallel parallelograms as opposite faces); XII applies the method of exhaustion introduced by Eudoxus to the volumes of solid figures, including the sphere; XIII, a three-dimensional analogue to Book IV, describes the Platonic solids. Among the jewels in Book XII is a proof of the recipe used by the Egyptians for the volume of a pyramid.

The Pythagorean theorem

Euclid's clever demonstration of the Pythagorean theorem in Book I became famous as the Windmill proof. The Pythagorean theorem states that the sum of the squares on the legs of a right triangle is equal to the square on the hypotenuse (the side opposite the right angle)—in familiar algebraic notation, $a^2 + b^2 = c^2$. The Babylonians and Egyptians had found some integer triples (a, b, c) satisfying the relationship. Pythagoras or one of his followers may

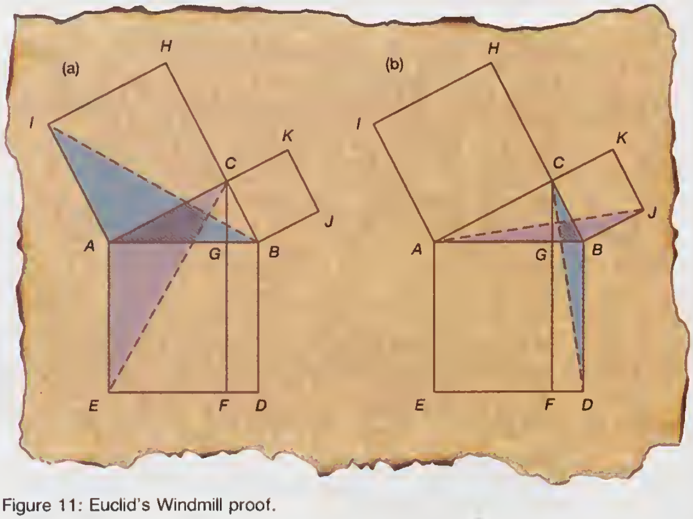


Figure 11: Euclid's Windmill proof.

have been the first to prove the theorem that bears his name. Euclid's proof, which generated the famous Windmill diagram (see Figure 11), is as follows:

1. Draw squares on the sides of the right ΔABC.
2. BCH and ACK are straight lines because $\angle ACB = 90°$
3. $\angle EAB = \angle CAI = 90°$, by construction.
4. $\angle BAI = \angle BAC + \angle CAI = \angle BAC + \angle EAB = \angle EAC$, by 3.
5. $AC = AI$ and $AB = AE$, by construction.
6. Therefore, $\Delta BAI \cong \Delta EAC$, by the side-angle-side theorem, as highlighted in part (a) of the figure.
7. Draw CF parallel to BD.
8. Rectangle $AGFE = 2\Delta ACE$. This remarkable result derives from two preliminary theorems: (a) the areas of all triangles on the same base, whose third vertex lies anywhere on an indefinitely extended line parallel to the base, are equal; and (b) the area of a triangle is half that of any parallelogram (including any rectangle) with the same base and height.
9. Square $AIHC = 2\Delta BAI$, by the same parallelogram theorem as in step 8.
10. Therefore, rectangle $AGFE$ = square $AIHC$, by steps 6, 8, and 9.
11. $\angle DBC = \angle ABJ$, as in steps 3 and 4.
12. $BC = BJ$ and $BD = AB$, by construction as in step 5.
13. $\Delta CBD \cong \Delta JBA$, as in step 6 and highlighted in part (b) of the figure.
14. Rectangle $BDFG = 2\Delta CBD$, as in step 8.
15. Square $CKJB = 2\Delta JBA$, as in step 9.
16. Thus, rectangle $BDFG$ = square $CKJB$, as in step 10.
17. Square $ABDE$ = rectangle $AGFE$ + rectangle $BDFG$, by construction.
18. Therefore, square $ABDE$ = square $AIHC$ + square $CKJB$, by steps 10 and 16.

GNOMONICS AND THE CONE

During its daily course above the horizon the Sun appears to describe a circular arc. Supplying in his mind's eye the missing portion of the daily circle, the Greek astronomer could imagine that his real eye was at the apex of a cone, the surface of which was defined by the Sun's rays at different times of the day and the base of which was defined by the Sun's apparent diurnal course. Our astronomer, using the pointer of a sundial, known as a gnomon, as his eye, would generate a second, shadow cone spreading downward. The intersection of this second cone with a horizontal surface, such as the face of a sundial, would give the trace of the Sun's image (or shadow) during the day as a plane section of a cone. (The possible intersections of a plane with a cone, known as the conic sections, are the circle, ellipse, point, straight line, parabola, and hyperbola; see Figure 12.)

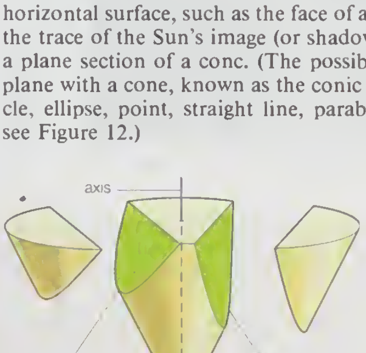


Figure 12: *Conic sections.*
The conic sections result from intersecting a plane with a double cone. There are three distinct families of conic sections: the ellipse (including the circle); the parabola (with one branch); and the hyperbola (with two branches).

However, the doxographers ascribe the discovery of conic sections to a student of Eudoxus, Menaechmus (mid-4th century BC), who used them to solve the problem of duplicating the cube. His restricted approach to conics—he worked with only right circular cones and made his sections at right angles to one of the straight lines composing their surfaces—was standard down to Archimedes' era. Euclid adopted Menaechmus's approach in his lost book on conics, and Archimedes followed suit. Doubtless, however, both knew that all the conics can be obtained from the same right cone by allowing the section at any angle.

The reason that Euclid's treatise on conics perished is that Apollonius of Perga (*c.* 262–*c.* 190 BC) did to it what Euclid had done to the geometry of Plato's time. Apollonius reproduced known results much more generally and discovered many new properties of the figures. He first proved that all conics are sections of any circular cone, right or oblique. Apollonius introduced the terms ellipse, hyperbola, and parabola for curves produced by intersecting a circular cone with a plane at an angle less than, greater than, and equal to, respectively, the opening angle of the cone.

ASTRONOMY AND TRIGONOMETRY

Calculation. In an inspired use of their geometry, the Greeks did what no earlier people seems to have done: they geometrized the heavens by supposing that the Sun, Moon, and planets move around a stationary Earth on a rotating circle or set of circles, and they calculated the speed of rotation of these supposititious circles from observed motions. Thus they assigned to the Sun a circle eccentric to the Earth to account for the unequal lengths of the seasons.

The Ptolemaic system

Ptolemy (fl. AD 127–145 in Alexandria, Egypt) worked out complete sets of circles for all the planets. In order to account for phenomena arising from the Earth's motion around the Sun, the Ptolemaic system included a secondary circle known as an epicycle, whose centre moved along the path of the primary orbital circle, known as the deferent. Ptolemy's *Great Compilation,* or *Almagest* after its Arabic translation, was to astronomy what Euclid's *Elements* was to geometry. Contrary to the *Elements,* however, the *Almagest* deploys geometry for the purpose of calculation. Among the items Ptolemy calculated was a table of chords, which correspond to the trigonometric sine function later introduced by Indian and Islāmic mathematicians. The table of chords assisted the calculation of distances from angular measurements as a modern astronomer might do with the law of sines.

Epistemology. The application of geometry to astronomy reframed the perennial Greek pursuit of the nature of truth. If a mathematical description fit the facts, as did Ptolemy's explanation of the unequal lengths of the seasons by the eccentricity of the Sun's orbit, should the description be taken as true of nature? The answer, with increasing emphasis, was "no." Astronomers remarked that the eccentric orbit representing the Sun's annual motion could be replaced by a pair of circles, a deferent centred on the Earth and an epicycle the centre of which moved along the circumference of the deferent. That gave two observationally equivalent solar theories based on two quite different mechanisms. Geometry was too prolific of alternatives to disclose the true principles of nature. The Greeks, who had raised a sublime science from a pile of practical recipes, discovered that in reversing the process, in reapplying their mathematics to the world, they had no securer claims to truth than the Egyptian rope pullers.

Ancient geometry: cosmological and metaphysical

PYTHAGOREAN NUMBERS AND PLATONIC SOLIDS

The Pythagoreans used geometrical figures to illustrate their slogan that all is number—thus their "triangular numbers" ($n(n-1)/2$), "square numbers" (n^2), and "altar numbers" (n^3), some of which are shown in Figure 13. This principle found a sophisticated application in Plato's creation story, the *Timaeus*, which presents the smallest particles, or "elements," of matter as regular geometrical figures. Since the ancients recognized four or five elements at most, Plato sought a small set of uniquely defined

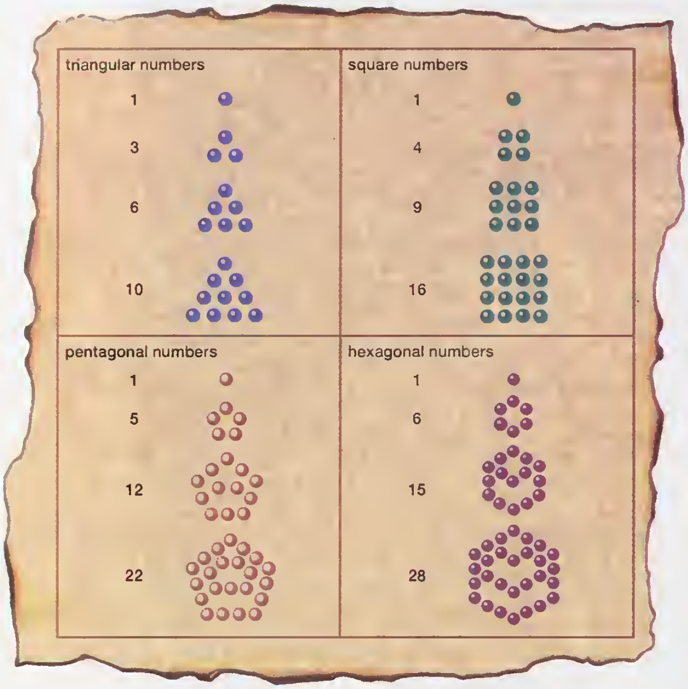


Figure 13: *Polygonal numbers.*
The ancient Greeks generally thought of numbers in concrete terms, particularly as measurements and geometric dimensions. Thus, they often arranged pebbles in various patterns to discern arithmetical, as well as mystical, relationships between the numbers. A few such patterns are indicated here.

geometrical objects to serve as elementary constituents. He found them in the only three-dimensional structures whose faces are equal regular polygons that meet one another at equal solid angles. As shown in Figure 14, they are: the tetrahedron, or pyramid (with 4 triangular faces); the cube (with 6 square faces); the octahedron (with 8 equilateral triangular faces); the dodecahedron (with 12 pentagonal faces); and the icosahedron (with 20 equilateral triangular faces).

The cosmology of the *Timaeus* had a consequence of the first importance for the development of mathematical astronomy. It guided Johannes Kepler (1571–1630) to his discovery of the laws of planetary motion. Kepler deployed the five regular Platonic solids not as indicators of the nature and number of the elements but as a model of the structure of the heavens. In 1596 he published *Prodromus Dissertationum Mathematicarum Continens Mysterium Cosmographicum* ("Cosmographic Mystery"), in which each of the known six planets revolved around the Sun on spheres separated by the five Platonic solids, as shown in Figure 15. Although Tycho Brahe (1546–1601), the world's greatest observational astronomer before the invention of the telescope, rejected the Copernican model of the solar system, he invited Kepler to assist him at his new observatory outside of Prague. In trying to resolve discrepancies between his original theory and Brahe's observations, Kepler made the capital discovery that the planets move in ellipses around the Sun as a focus.

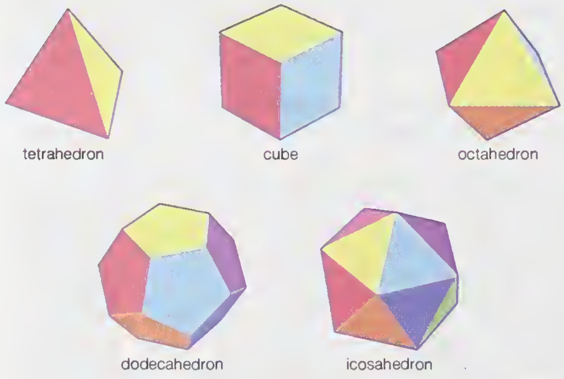


Figure 14: The five Platonic solids.

MEASURING THE EARTH AND THE HEAVENS

Eratosthenes' measurement. Geometry offered Greek cosmologists not only a way to speculate about the structure of the universe but also the means to measure it. South of Alexandria and roughly on the same meridian of longitude is the village of Syene (modern Aswān), where the Sun stands directly overhead at noon on a midsummer day. At the same moment at Alexandria, the Sun's rays make an angle α with the tip of a vertical rod, as shown in Figure 16. Since the Sun's rays fall almost parallel on the Earth, the angle subtended by the arc l (representing the distance between Alexandria and Syene) at the centre of the Earth also equals α; thus the ratio of the Earth's circumference, C, to the distance, l, must equal the ratio of 360° to the angle α—in symbols, $C{:}l = 360°{:}\alpha$. Eratosthenes made the measurements, obtaining a value of about 5,000 stadia for l, which gave a value for the Earth's circumference of about 400,000 stadia. Because the accepted length of the Greek stadium varied locally, we cannot accurately determine Eratosthenes' margin of error. However, if we credit the ancient historian Plutarch's guess at Eratosthenes' unit of length, we obtain a value for the Earth's circumference of about 46,250 kilometres (27,750 miles)—about 15 percent larger than the modern value but remarkably close, considering the difficulty in accurately measuring l and α.

By the permission of the British Library

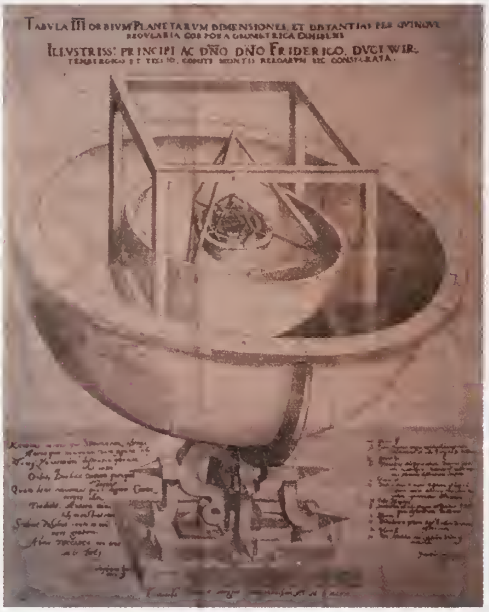

TABVLA III ORBIVM PLANETARVM DIMENSIONES, ET DISTANTIAS PER QVINQVE
ILLVSTRISS: PRINCIPI AC DÑO DÑO FRIDERICO, DVCI WIR-

Figure 15: Illustration of the cosmos in Johannes Kepler, *Mysterium Cosmographicum* (1596).

Poseidonius's measurement. In addition to the attempts of Eratosthenes to measure the Earth, two other early attempts had a lasting historical impact, since they provided values that Christopher Columbus (1451–1506) exploited in selling his project to reach Asia by traveling west from Europe. One was devised by the Greek philosopher Poseidonius (*c.* 135–*c.* 51 BC), the teacher of the great Roman statesman Marcus Tullius Cicero (106–43 BC). According to Poseidonius, when the star Canopus sets at Rhodes, it appears to be 7.5° above the horizon at Alexandria. (In fact, it is a little over 5°.) The situation appears in Figure 17, where the dark lines represent the horizons at Rhodes (R) and Alexandria (A). Because of the right angles at R and A and the parallel lines of sight to Canopus, $\angle RCA$ equals the angular height of Canopus at Alexandria (the errant 7.5°). To obtain the radius $r = CR = CA$, Poseidonius needed the length of the arc RA. It could not be paced out, as travelers from Aswān to Alexandria had done for Eratosthenes' result, because the journey lay over water. Poseidonius could only guess the distance, and his calculation

Christopher Columbus selling his voyage

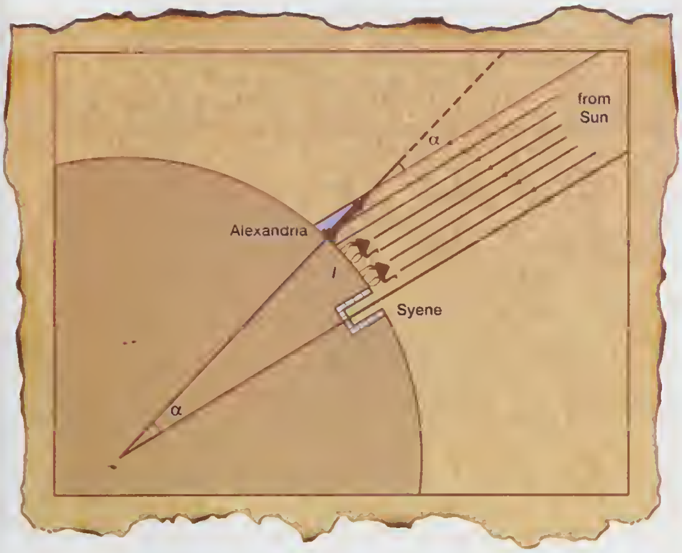


Figure 16: Eratosthenes' measurement of the Earth.

for the size of the Earth was less than three-quarters of what Eratosthenes had found.

Arabic measurement. The second method, devised by the Muslim mathematician Abu al-Bīrūnī (973–1048), required a free-standing mountain of known height AB (see Figure 18). The observer measured $\angle ABH$ between the vertical BA and the line to the horizon BH. Since $\angle BHC$ is a right angle, the Earth's radius $r = CH = AC$ is given by solution of the simple trigonometric equation $\sin(\angle ABH) = r/(r + AB)$. The Arab value for the Earth's circumference agreed with the value calculated by Poseidonius—or so Columbus argued, ignoring or forgetting that the Arabs expressed their results in Arab miles, which were longer than the Roman miles with which Poseidonius worked. By claiming that the "best" measurements agreed that the real Earth was three-fourths the size of Eratosthenes' Earth, Columbus reassured his backers that his small wooden ships could survive the journey—he put it at 30 days—to "Cipangu" (Japan).

Measuring the solar distance. Aristarchus of Samos (*c.* 310–230 BC) has garnered the credit for extending the grip of number as far as the Sun. Using the Moon as a ruler and noting that the apparent sizes of the Sun and the Moon are about equal, he calculated values for his treatise "On the Sizes and Distances of the Sun and Moon." The great difficulty of making the observations resulted in an underestimation of the solar distance about 20-fold—he obtained a solar distance roughly 1,200 times the Earth's radius. Possibly Aristarchus's inquiry into the relative sizes of the Sun, Moon, and Earth led him to propound the first heliocentric ("Sun-centred") model of the universe.

Aristarchus's value for the solar distance, no matter how mistaken, was confirmed by an astonishing coincidence. Ptolemy equated the maximum distance of the Moon in its eccentric orbit with the closest approach of Mercury riding on its epicycle; the farthest distance of Mercury with the closest of Venus; and the farthest of Venus with the closest of the Sun. Thus he could compute the solar distance in terms of the lunar distance and thence the terrestrial radius. His answer agreed with that of Aristarchus. The Ptolemaic conception of the order and machinery of the planets, the most powerful application of Greek geometry to the physical world, thus corroborated the result of direct measurement and established the dimensions of the cosmos for over a thousand years. As the ancient philosophers said, there is no truth in astronomy.

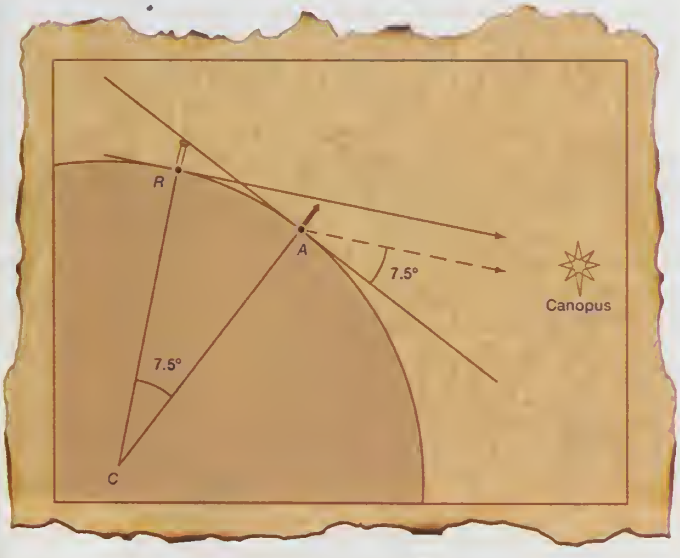


Figure 17: Poseidonius measures the Earth *c.* 100 BC.

The postclassical period

PASSAGE THROUGH ISLĀM

Two centuries after they broke out of their desert around Mecca, the followers of Muḥammad occupied the lands from Persia to Spain and settled down to master the arts and sciences of the peoples they had conquered. They admired especially the works of the Greek mathematicians and physicians and the philosophy of Aristotle. By the late 9th century they were already able to add to the geometry of Euclid, Archimedes, and Apollonius. In the 10th century they went beyond Ptolemy. Stimulated by the problem of finding the effective orientation for prayer (the *qibla,* or direction from the place of worship to Mecca), Islāmic geometers and astronomers developed the stereographic projection (invented to project the celestial sphere onto a two-dimensional map or instrument), as well as plane and spherical trigonometry. Here they incorporated elements derived from India as well as from Greece. Their achievements in geometry and geometrical astronomy materialized in instruments for drawing conic sections and, above all, in the beautiful brass astrolabes with which they reduced to the turn of a dial the toil of calculating astronomical quantities.

Finding the *qibla*

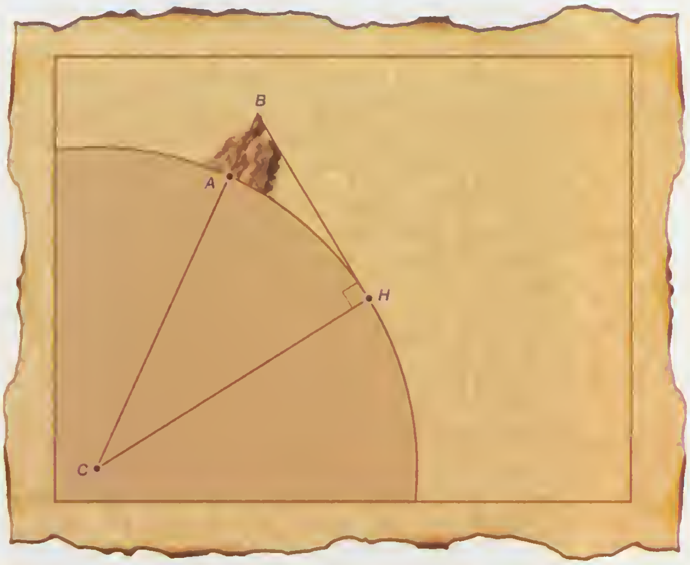


Figure 18: Arabic method of measuring the Earth.

Thābit ibn Qurrah (836–901) had precisely the attributes required to bring the geometry of the Arabs up to the mark set by the Greeks. As a member of a religious sect close but hostile to both Jews and Christians, he knew Syriac and Greek as well as Arabic; as a money changer, he knew how to calculate; as both, he recommended himself to the Banū Mūsā, a set of mathematician brothers descended from a robber who had diversified into astrology. The Banū Mūsā

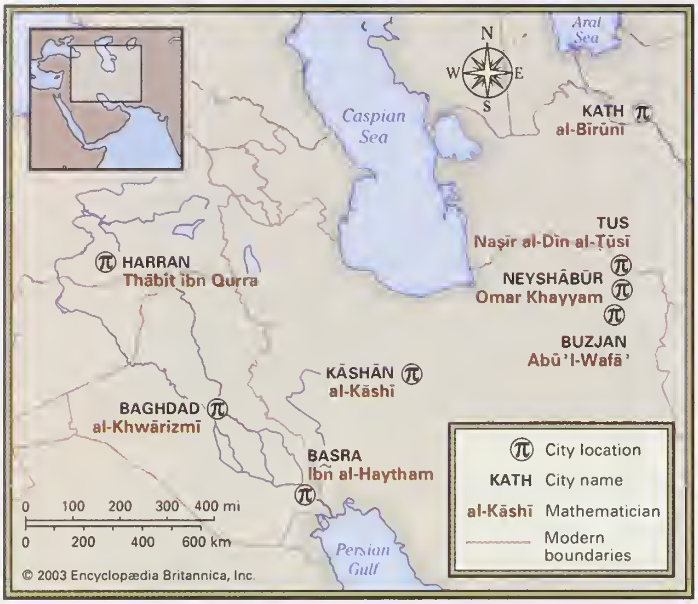


Figure 19: *Mathematicians of the Islāmic world.*
This map spans more than 600 years of prominent Islāmic mathematicians, from al-Khwārizmī (*c.* AD 600) to al-Kāshī (c. AD 1400).

House of Wisdom

directed a House of Wisdom in Baghdad sponsored by the caliph. There they presided over translations of the Greek classics. Thābit became an ornament of the House of Wisdom. He translated Archimedes and Apollonius, some of whose books now are known only in his versions. In a notable addition to Euclid, he tried valiantly to prove the parallel postulate (see below, *Non-Euclidean geometries*).

Among the pieces of Greek geometrical astronomy that the Arabs made their own was the planispheric astrolabe, which incorporated one of the methods of projecting the celestial sphere onto a two-dimensional surface invented in ancient Greece. One of the desirable mathematical features of this method (the stereographic projection) is that it converts circles into circles or straight lines, a property proved in the first pages of Apollonius's *Conics*. As Ptolemy showed in his *Planisphaerium*, the fact that the stereographic projection maps circles into circles or straight lines makes the astrolabe a very convenient instrument for reckoning time and representing the motions of celestial bodies. The earliest known Arabic astrolabes and manuals for their construction date from the 9th century. The Islāmic world improved the astrolabe as an aid for determining the time for prayers, for finding the direction to Mecca, and for astrological divination.

EUROPE REDISCOVERS THE CLASSICS

Contacts among Christians, Jews, and Arabs in Catalonia brought knowledge of the astrolabe to the West before the year 1000. During the 12th century many manuals for its use and construction were translated into Latin along with geometrical works by the Banū Mūsā, Thābit, and others. Some of the achievements of the Arab geometers were rediscovered in the West after wide and close study of Euclid's *Elements*, which was translated repeatedly from the Arabic and once from the Greek in the 12th and 13th centuries. The *Elements* (Venice, 1482) was one of the first technical books ever printed. Archimedes also came West in the 12th century, in Latin translations from Greek and Arabic sources. Apollonius arrived only by bits and pieces. Ptolemy's *Almagest* appeared in Latin manuscript in 1175. Not until the humanists of the Renaissance turned their classical learning to mathematics, however, did the Greeks come out in standard printed editions in both Latin and Greek.

These texts affected their Latin readers with the strength of revelation. Europeans discovered the notion of proof, the power of generalization, and the superhuman cleverness of the Greeks; they hurried to master techniques that would enable them to improve their calendars and horoscopes, fashion better instruments, and raise Christian mathematicians to the level of the infidels. It took more than two centuries for the Europeans to make their unexpected heritage their own. By the 15th century, however, they were prepared to go beyond their sources. The most novel developments occurred where creativity was strongest, in the art of the Italian Renaissance.

LINEAR PERSPECTIVE

The theory of linear perspective, the brainchild of the Florentine architect-engineers Filippo Brunelleschi (1377–1446) and Leon Battista Alberti (1404–72) and their followers, was to help remake geometry during the 17th century. The scheme of Brunelleschi and Alberti, as given without proofs in Alberti's *De pictura* (1435; *On Painting*), exploits the pyramid of rays that, according to what they had learned from the Westernized versions of the optics of Ibn Al-Haytham (*c.* 965–1040), proceeds from the object to the painter's eye. Imagine, as Alberti directed, that the painter studies a scene through a window, using only one eye and not moving his head; he cannot know whether he looks at an external scene or at a glass painted to present to his eye the same visual pyramid. Supposing this decorated window to be the canvas, Alberti interpreted the painting-to-be as the projection of the scene in life onto a vertical plane cutting the visual pyramid. A distinctive feature of his system was the "point at infinity" at which parallel lines in the painting appear to converge, as shown in Figure 20.

Point of infinity

Alberti's procedure, as developed by Piero della Francesca (*c.* 1410–92) and Albrecht Dürer (1471–1528), was used by many artists who wished to render perspective persuasively. At the same time, cartographers tried various projections of the sphere to accommodate the record of geographical discoveries that began in the mid-15th century with Portuguese exploration of the west coast of Africa. Coincidentally with these explorations, mapmakers recovered Ptolemy's *Geography*, in which he had recorded by latitude (sometimes near enough) and longitude (usually

Scala/Art Resource

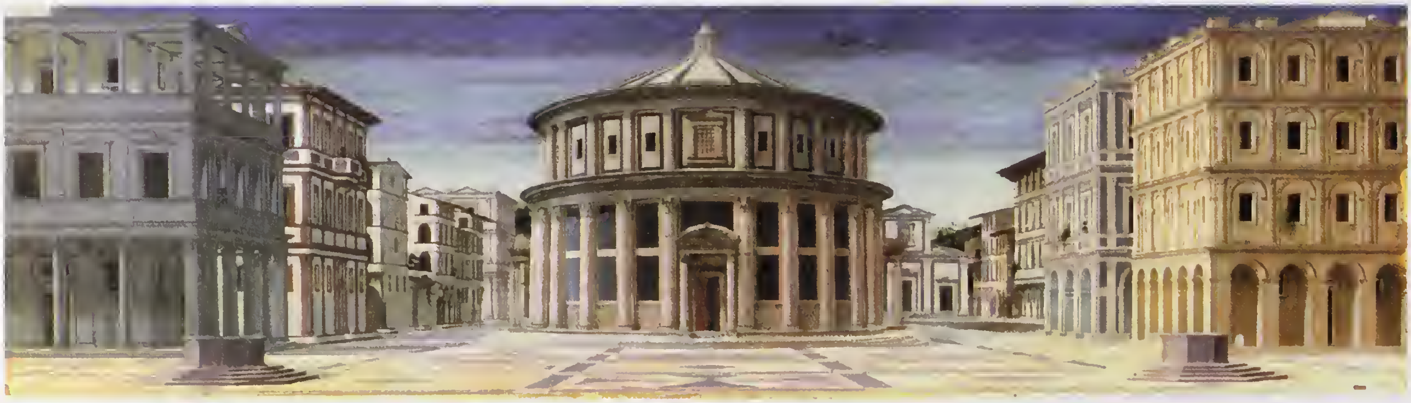


Figure 20: Piero della Francesca's *Ideal City* (c. 1470), in the Galleria Nazionale delle Marche, Urbino, Italy.

Figure 21: World map after Ptolemy, *Geographia* (Ulm, 1496), Bibliotheque Nationale, Paris, France.
Giraudon/Art Resource

far off) the principal places known to him and indicated how they could be projected onto a map.

The discoveries that enlarged the known Earth did not fit easily on Ptolemy's projections. Cartographers therefore adopted the stereographic projection that had served astronomers. Several projected the Northern Hemisphere onto the Equator just as in the standard astrolabe, but the most widely used aspect, popularized in the world maps made by Gerardus Mercator's son for later editions of his father's atlas (beginning in 1595), projected points on the Earth onto a cylinder tangent to the Earth at the Equator. After cutting the cylinder along a vertical line and flattening the resulting rectangle, the result was the now-familiar Mercator projection map.

By permission of the British Library

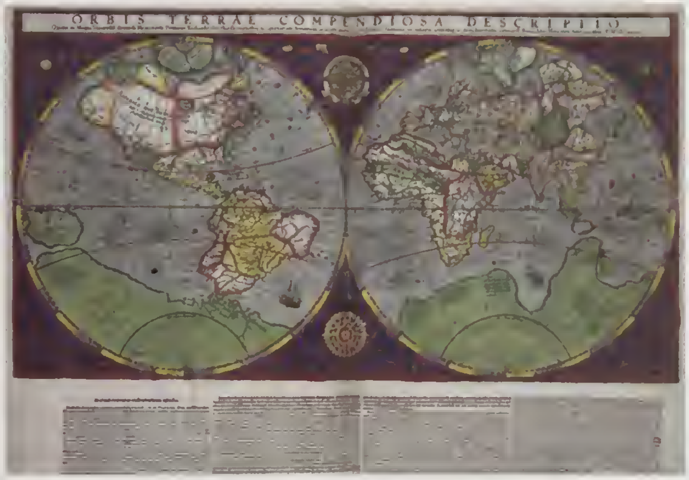


Figure 22: Mercator projection map, *Orbis terrae compendiosa descriptio*, 1587.

The intense cultivation of methods of projection by artists, architects, and cartographers during the Renaissance eventually provoked mathematicians into considering the properties of linear perspective in general. The most profound of these generalists was a sometime architect named Girard Desargues (1591–1661).

Transformation

FRENCH CIRCLES

Desargues was a member of intersecting circles of 17th-century French mathematicians worthy of Plato's Academy of the 4th century BC or Baghdad's House of Wisdom of the 9th century AD. They included René Descartes (1596–1650) and Pierre de Fermat (1601–65), inventors of analytic geometry; Gilles Personne de Roberval (1602–75), a pioneer in the development of the calculus; and Blaise Pascal (1623–62), a contributor to the calculus and an exponent of the principles set forth by Desargues.

Projective geometry. Two main directions can be distinguished in Desargues's work. Like Renaissance artists, Desargues freely admitted the point at infinity into his demonstrations and showed that every set of parallel lines in a scene (apart from those parallel to the sides of the canvas) should project as converging bundles at some point on the "line at infinity" (the horizon). With the addition of points at infinity to the Euclidean plane, Desargues could frame all his propositions about straight lines without excepting parallel ones—which, like the others, now met one another, although not before "infinity." A farther-reaching matter arising from artistic perspective was the relation between projections of the same object from different points of view and different positions of the canvas. Desargues observed that neither size nor shape is generally preserved in projections, but collinearity is, and he provided an example, possibly useful to artists, in images of triangles seen from different points of view. The statement that accompanied this example became known as Desargues's theorem.

Desargues's theorem

Desargues's second direction was to "simplify" Apollonius's work on conic sections. Despite his generality of approach, Apollonius had to prove all his theorems for each type of conic separately. Desargues saw that he could prove them all at once and, moreover, by treating a cylinder as a cone with vertex at infinity, demonstrate useful analogies between cylinders and cones. Following his lead, Pascal made his surprising discovery that the intersections of the three pairs of opposite sides of a hexagon inscribed in a conic lie on a straight line. (See Figure 23.) In 1685, in his *Sectiones Conicæ*, Philippe de la Hire (1640–1718), a Parisian painter turned mathematician, proved several hundred propositions in Apollonius's *Conics* by Desargues's efficient methods.

Cartesian geometry. In 1619, as part of the great illumination that inspired Descartes to assume the modest chore of reforming philosophy as well as mathematics, he devised "compasses" made of sticks sliding in grooved frames to duplicate the cube and trisect angles. Descartes esteemed these implements and the constructions they effected as (to quote from a letter of 1619) "no less certain and geometrical than the ordinary ones with which circles are drawn." By the use of apt instruments, he would bring ancient mathematics to perfection: "scarcely anything will remain to be discovered in geometry."

What Descartes had in mind was the use of compasses with sliding members to generate curves. To classify and

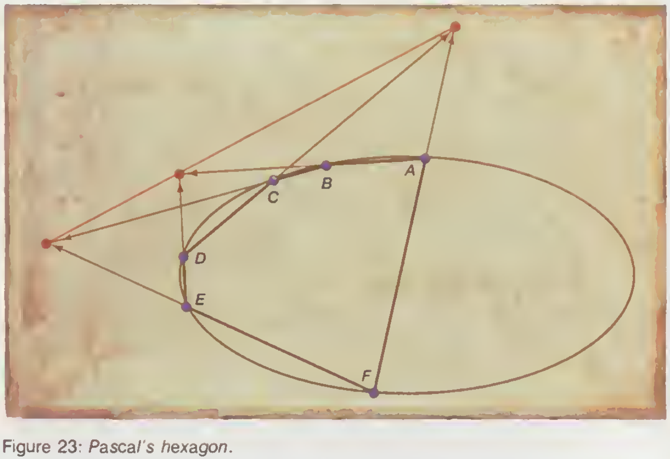


Figure 23: *Pascal's hexagon.*
Blaise Pascal proved that for any hexagon inscribed in any conic section (ellipse, parabola, hyperbola) the three pairs of opposite sides when extended intersect in points that lie on a straight line. In the figure an irregular hexagon is inscribed in an ellipse. Opposite sides *AB* and *DE*, *BC* and *EF*, and *CD* and *AF* intersect at points on a line outside the ellipse.

study such curves, Descartes took his lead from the relations Apollonius had used to classify conic sections, which contain the squares, but no higher powers, of the variables. To describe the more complicated curves produced by his instruments or defined as the loci of points satisfying involved criteria, Descartes had to include cubes and higher powers of the variables. He thus overcame what he called the deceptive character of the terms square, rectangle, and cube as used by the ancients and came to identify geometric curves as depictions of relationships defined algebraically. By reducing relations difficult to state and prove geometrically to algebraic relations between coordinates (usually rectangular) of points on curves, Descartes brought about the union of algebra and geometry that gave birth to the calculus.

GEOMETRICAL CALCULUS

The familiar use of infinity, which underlay much of perspective theory and projective geometry, also leavened the tedious Archimedean method of exhaustion. Not surprisingly, a practical man, the Flemish engineer Simon Stevin (1548–1620), who wrote on perspective and cartography among many other topics of applied mathematics, gave the first effective impulse toward redefining the object of Archimedean analysis. Instead of confining the circle between an inscribed and a circumscribed polygon, the new view regarded the circle as identical to the polygons, and the polygons to one another, when the number of their sides becomes infinitely great.

Cavalieri's indivisibles

Cavalieri's method. This revitalized approach to exhaustion received a preliminary systematization in the *Geometria Indivisibilibus Continuorum Nova Quadam Ratione Promota* (1635; "A Method for the Determination of a New Geometry of Continuous Indivisibles") by the Italian mathematician Bonaventura (Francesco) Cavalieri (1598–1647). Cavalieri, perhaps influenced by Kepler's method of determining volumes in *Nova Steriometria Doliorum* (1615; "New Stereometry of Wine Barrels"), regarded lines as made up of an infinite number of dimensionless points, areas as made up of lines of infinitesimal thickness, and volumes as made up of planes of infinitesimal depth in order to obtain algebraic ways of summing the elements into which he divided his figures. Cavalieri's method may be stated as follows: if two figures (solids) of equal height are cut by parallel lines (planes) such that each pair of lengths (areas) matches, then the two figures (solids) have the same area (volume). (See Figure 24.) Although not up to the rigorous standards of today and criticized by "classicist" contemporaries (who were unaware that Archimedes himself had explored similar techniques), Cavalieri's method of indivisibles became a standard tool for solving volumes until the introduction of integral calculus near the end of the 17th century.

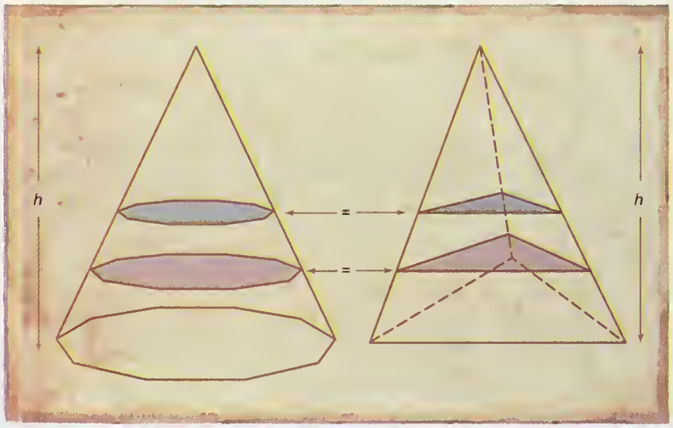


Figure 24: *Cavalieri's principle.*
Take a regular polygon equal in area to an equilateral triangle. Erect a pyramid on the triangle and a cone-like figure of the same height on the polygon. Cross sections of both figures taken at the same height above the bases are equal. Therefore, by Cavalieri's theorem, so are the volumes of the solids.

Fermat's method. A second geometrical inspiration for the calculus derived from efforts to define tangents to curves more complicated than conics. Fermat's method, representative of many, had as its exemplar the problem of finding the rectangle that maximizes the area for a given perimeter. Let the sides sought for the rectangle be denoted by a and b. Increase one side and diminish the other by a small amount ε; the resultant area is then given by $(a + \varepsilon)(b - \varepsilon)$. Fermat observed what Kepler had perceived earlier in investigating the most useful shapes for wine casks, that near its maximum (or minimum) a quantity scarcely changes as the variables on which it depends alter slightly. On this principle, Fermat equated the areas ab and $(a + \varepsilon)(b - \varepsilon)$ to obtain the stationary values: $ab = ab - \varepsilon a + \varepsilon b - \varepsilon^2$. By canceling the common term ab, dividing by ε, and then setting ε at zero, Fermat had his well-known answer, $a = b$. The figure with maximum area is a square.

To obtain the tangent to a curve at a point P (x, y), Fermat began by drawing a secant line to a nearby point P_1 $(x + \varepsilon, y_1)$, as shown in Figure 25. Thus, for small ε, the secant line PP_1 is approximately equal to the true tangent line PT. Finally, Fermat allowed ε to shrink to zero, thus obtaining a mathematical expression for the true tangent line.

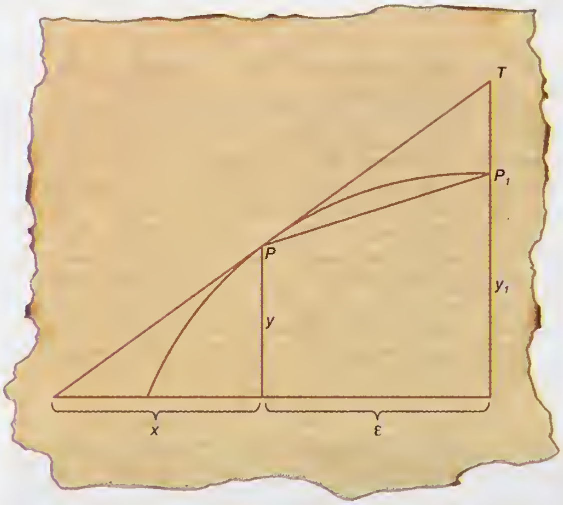


Figure 25: Fermat's tangent method.

THE WORLD SYSTEM

Part of the motivation for the close study of Apollonius during the 17th century was the application of conic sections to astronomy. Kepler not only replaced the many circles of the old planetary system with a few ellipses, he also substituted a complicated rule of motion (his "second law") for the relatively simple Ptolemaic rule that all motions must be compounded of rotations performed at constant velocity. Kepler's second law states that a planet moves in its ellipse so that the line between it and the Sun placed at a focus sweeps out equal areas in equal times. His astronomy thus made pressing and practical the otherwise merely difficult problem of the quadrature of conics and the associated theory of indivisibles.

Newton and planetary dynamics

With the methods of Apollonius and a few infinitesimals, an inspired geometer showed that the laws regarding both area and ellipse can be derived from the suppositions that bodies free from all forces either rest or travel uniformly in straight lines and that each planet constantly falls toward the Sun with an acceleration that depends only on the distance between their centres. The inspired geometer was Isaac Newton (1642 [Old Style]–1727), who made planetary dynamics a matter entirely of geometry by replacing the planetary orbit by a succession of infinitesimal chords, planetary acceleration by a series of centripetal jerks, and, in keeping with Kepler's second law, time by an area.

Besides the problem of planetary motion, questions in optics pushed 17th-century natural philosophers and mathematicians to the study of conic sections. As Archimedes is supposed to have shown (or shone) in his destruction of a Roman fleet by reflected sunlight, a parabolic mirror brings all rays parallel to its axis to a common focus. The story of Archimedes provoked many later geometers, including Newton, to emulation. Eventually they created instruments powerful enough to melt iron.

The fitting of lenses to surveying instruments in the 1660s greatly improved the accuracy of the Greek method of measuring the Earth, and this soon became the preferred technique. In its modern form, the method requires the following elements: two stations on the same meridian of longitude, which play the same parts as Aswān and Alexandria in the method of Eratosthenes; a precise determination of the angular height of a designated star at the same time from the two stations; and two perfectly level and accurately measured baselines a few kilometres long near each station. What was new 2,000 years after Eratosthenes was the accuracy of the stellar positions and the measured distance between the stations, accomplished through the use of the baselines. At each end of one baseline surveyors raise tall posts that can be seen from some nearby vantage point, say a church steeple, and the angle between the posts is measured. From a second viewpoint, say the top of a tree, the angle made between one of the posts and the steeple is taken. Observation from a third station gives an angle between the treetop and the steeple. Proceeding thus from positions on either side of the line to be measured, the surveyors create a series of virtual triangles whose sides they can compute trigonometrically from the observed angles and the measured length of the first baseline. The closeness of agreement between the calculation based on the first baseline and the measurement of the second baseline gives a check on the work.

During the 18th century surveyors and astronomers, practicing their updated Greek geodesy in Lapland and Peru, corroborated the conclusion of Newton, deduced at his desk in Cambridge, England, that the Earth's equatorial axis exceeds its polar axis by a few miles. So precise was the method that subsequent investigation using it revealed that the Earth does not have the shape of an ellipsoid of revolution (an ellipse rotated around one of its axes) but rather has an ineffable shape of its own, now known as the geoid. The method further established the fundamental grids for the mapping of Europe and its colonies. During the French Revolution modernized Greek geodesy was employed to find the equivalent, in the old royal system of measurement, of the new fundamental unit, the standard metre. By definition, the metre was one ten-millionth part of a quarter of the meridian through Paris, making the Earth circumference a nominal 40,000 kilometres.

Relaxation and rigour

The dominance of analysis (algebra and the calculus) during the 18th century produced a reaction in favour of geometry early in the 19th century. Fundamental new branches of the subject resulted that deepened, generalized, and violated principles of ancient geometry. The cultivators of these new fields, such as Jean-Victor Poncelet (1788–1867) and his self-taught disciple Jakob Steiner (1796–1863), vehemently urged the claims of geometry over analysis. The early 19th-century revival of pure geometry produced the discovery that Euclid had devoted his efforts to only one of several comprehensive geometries, the others of which can be created by replacing Euclid's fifth postulate with another about parallels.

PROJECTION AGAIN

Poncelet, who was an officer in the French corps of engineers, learned scraps of Desargues's work from his teacher Gaspard Monge (1746–1818), who developed his own method of projection for drawings of buildings and machines. Poncelet relied on this information to keep himself alive. Taken captive during Napoleon's invasion of Russia in 1812, he passed his time by rehearsing in his head the things he had learned from Monge. The result was projective geometry.

Poncelet employed three basic tools. One he took from Desargues's theorem: the demonstration of difficult theorems about a complicated figure by working out equivalent simpler theorems on an elementary figure interchangeable with the original figure by projection. The second tool, continuity, allows the geometer to claim certain things as true for one figure that are true of another equally general figure provided that the figures can be derived from one another by a certain process of continual change. Poncelet and his defender Michel Chasles (1793–1880) extended the principle of continuity into the domain of the imagination by considering constructs such as the common chord in two circles that do not intersect.

Principle of duality

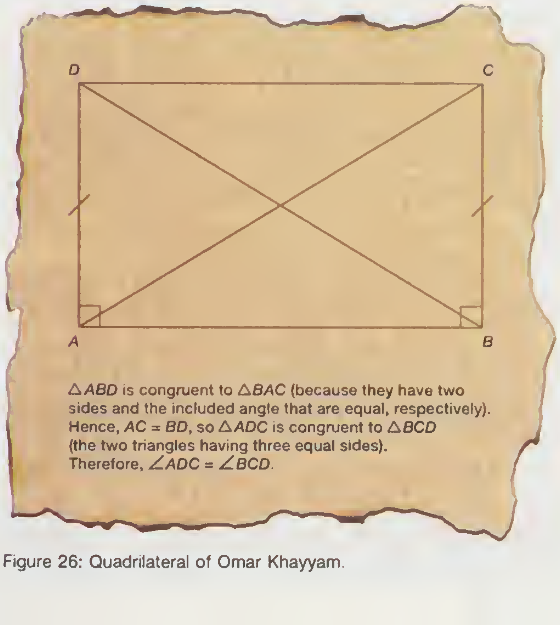
Poncelet's third tool was the "principle of duality," which interchanges various concepts such as points with lines, or lines with planes, so as to generate new theorems from old theorems. Desargues's theorem allows their interchange. So, as Steiner showed, does Pascal's theorem that the three points of intersection of the opposite sides of a hexagon inscribed in a conic lie on a line; thus, the lines joining the opposite vertices of a hexagon circumscribed about a conic meet in a point. (See Figure 23.)

Poncelet's followers realized that they were hampering themselves, and disguising the true fundamentality of projective geometry, by retaining the concept of length and congruence in their formulations, since projections do not usually preserve them. Similarly, parallelism had to go. Efforts were well under way by the middle of the 19th century, by Karl George Christian von Staudt (1798–1867) among others, to purge projective geometry of the last superfluous relics from its Euclidean past.

NON-EUCLIDEAN GEOMETRIES

The Enlightenment was not so preoccupied with analysis as to completely ignore the problem of Euclid's fifth postulate. In 1733 Girolamo Saccheri (1667–1733), a Jesuit professor of mathematics at the University of Pavia, Italy, substantially advanced the age-old discussion by setting forth the alternatives in great clarity and detail before declaring that he had "cleared Euclid of every defect" (*Euclides ab Omni Naevo Vindicatus,* 1733). Euclid's fifth postulate runs: "If a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the straight lines, if produced indefinitely, will meet on that side on which are the angles less than two right angles." Saccheri took up the quadrilateral of Omar Khayyam (1048–1131), who started with two parallel lines AB and DC, formed the sides by drawing lines AD and BC perpendicular to AB, and then considered three hypotheses for the internal angles at C and D: to be right, obtuse, or acute (see Figure 26). The first possibility gives Euclidean geometry. Saccheri devoted himself to proving that the obtuse and the acute alternatives both end in contradictions, which would thereby eliminate the need for an explicit parallel postulate.

On the way to this spurious demonstration, Saccheri established several theorems of non-Euclidean geometry—for example, that according to whether the right, obtuse, or



$\triangle ABD$ is congruent to $\triangle BAC$ (because they have two sides and the included angle that are equal, respectively). Hence, $AC = BD$, so $\triangle ADC$ is congruent to $\triangle BCD$ (the two triangles having three equal sides). Therefore, $\angle ADC = \angle BCD$.

Figure 26: Quadrilateral of Omar Khayyam.

acute hypothesis is true, the sum of the angles of a triangle respectively equals, exceeds, or falls short of 180°. He then destroyed the obtuse hypothesis by an argument that depended upon allowing lines to increase in length indefinitely. If this is disallowed, the hypothesis of the obtuse angle produces a valid system that is equivalent to standard spherical geometry, the geometry of figures drawn on the surface of a sphere.

As for the acute angle, Saccheri could defeat it only by appealing to an arbitrary hypothesis about the behaviour of lines at infinity. One of his followers, the Swiss-German polymath Johann Heinrich Lambert (1728–77), observed that, based on the acute hypothesis, the area of a triangle is the negative of that of a spherical triangle. Since the latter is proportional to the square of the radius, r, the former appeared to Lambert to be the area of an imaginary sphere with radius ir, where $i = \sqrt{-1}$, which he deemed ridiculous.

Although both Saccheri and Lambert aimed to establish the hypothesis of the right angle, their arguments seemed rather to indicate the unimpeachability of the alternatives. Several mathematicians at the University of Göttingen, notably the great Carl Friedrich Gauss (1777–1855), then took up the problem. Gauss was probably the first to perceive that a consistent geometry could be built up independent of Euclid's fifth postulate, and he derived many relevant propositions, which, however, he promulgated only in his teaching and correspondence. The earliest published non-Euclidean geometric systems were the independent work of two young men from the East who had nothing to lose by their boldness. Both can be considered Gauss's disciples once removed: the Russian Nikolay Ivanovich Lobachevsky (1792–1856), who learned his mathematics from a close friend of Gauss's at the University of Kazan, where Lobachevsky later became a professor; and János Bolyai (1802–60), an officer in the Austro-Hungarian army whose father also was a friend of Gauss's. Both Lobachevsky and Bolyai had worked out their novel geometries by 1826.

Lobachevsky and Bolyai reasoned about the hypothesis of the acute angle in the manner of Saccheri and Lambert and recovered their results about the areas of triangles. They advanced beyond Saccheri and Lambert by deriving an imaginary trigonometry to go with their imaginary geometry. Just as Desargues's projective geometry was neglected for many years, so the work of Bolyai and Lobachevsky made little impression on mathematicians for a generation and more. It was largely the posthumous publication in 1855 of Gauss's ideas about non-Euclidean geometry that gave the new approaches the cachet to attract the attention of later mathematicians.

A GRAND SYNTHESIS

Another of the profound impulses Gauss gave geometry concerned the general description of surfaces. Typically—with the notable exception of the geometry of the sphere—mathematicians had treated surfaces as structures in three-dimensional Euclidean space. However, as these surfaces occupy only two dimensions, only two variables are needed to describe them. This prompted the thought that two-dimensional surfaces could be considered as "spaces" with their own geometries, not just as Euclidean structures in ordinary space. For example, the shortest distance, or path, between two points on the surface of a sphere is the lesser arc of the great circle joining them, whereas, considered as points in three-dimensional space, the shortest distance between them is an ordinary straight line.

The shortest path between two points on a surface lying wholly within that surface is called a geodesic, which reflects the origin of the concept in geodesy, in which Gauss took an active interest. His initiative in the study of surfaces as spaces and geodesics as their "lines" was pursued by his student and, briefly, his successor at Göttingen, Bernhard Riemann (1826–66). Riemann began with an abstract space of n dimensions. That was in the 1850s, when mathematicians and mathematical physicists were beginning to use n-dimensional Euclidean space to describe the motions of systems of particles in the then-new kinetic theory of gases. Riemann worked in a quasi-Euclidean space—"quasi" because he used the calculus to

Abstract n-dimensional Euclidean space

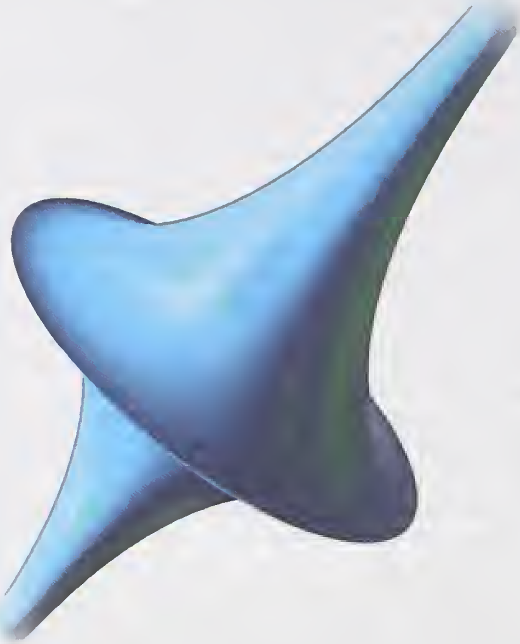


Figure 27: The pseudosphere.

generalize the Pythagorean theorem to supply sufficient flexibility to provide for geodesics on any surface.

When this very general differential geometry came down to two-dimensional surfaces of constant curvature, it revealed excellent models for non-Euclidean geometries. Riemann himself pointed out that, merely by calling the geodesics of a sphere "straight lines," the maligned hypothesis of the obtuse angle produces the geometry appropriate to the sphere's surface. Similarly, as shown by Eugenio Beltrami (1835–1900), who ended his teaching career in Saccheri's old post at Pavia, the geometry defined in the plane by the hypothesis of the acute angle fits perfectly a surface of revolution of constant negative curvature now called a pseudosphere (see Figure 27)—again, provided that its geodesics are accepted as the straight lines of the geometry.

Since the hypothesis of the obtuse angle correctly characterizes Euclidean geometry applied to the surface of a sphere, the non-Euclidean geometry based on it must be exactly as consistent as Euclidean geometry. The case of the acute angle treated by Lobachevsky and Bolyai required a sharper tool. Beltrami found it in a projection into a disc in the Euclidean plane of the points of a non-Euclidean space, in which each geodesic from the non-Euclidean space corresponds to a chord of the disc (see Figure 49 in *Hyperbolic Geometry,* below). Geometry built on the hypothesis of the acute angle has the same consistency as Euclidean geometry.

The key role of Euclidean geometry in proofs of the consistency of non-Euclidean geometries exposed the *Elements* to ever-deeper scrutiny. The old blemishes—particularly appeals to intuition and diagrams for the meaning of concepts like "inside" and "between" and the use of questionable procedures like superposition to prove congruency—became intolerable to mathematicians who laboured to clarify the foundations of arithmetic and the calculus as well as the interrelations of the new geometries. The German mathematician Moritz Pasch (1843–1930), in his *Vorlesungen über neuere Geometrie* (1882; "Lectures on the New Geometry"), identified what was wanting: undefined concepts, axioms about those concepts, and more rigorous logic based on those axioms. The choice of undefined concepts and axioms is free, apart from the constraint of consistency. Mathematicians following Pasch's path introduced various elements and axioms and developed their geometries with greater or lesser elegance and trouble. The most successful of these systematizers was the Göttingen professor David Hilbert (1862–1943), whose

Foundations of Geometry (1899) greatly influenced efforts to axiomatize all of mathematics.

THE REAL WORLD

Euclid's *Elements* had claimed to be a true account of space. Within this interpretation, Euclid's fifth postulate was an empirical finding; non-Euclidean geometries did not apply to the real world. Bolyai apparently could not free himself from the persuasion that Euclidean geometry represented reality. Lobachevsky observed that, if there were a star so distant that its parallax was not observable from the Earth's orbit, his geometry would be indistinguishable from Euclid's at the point where the parallax vanished. By his calculation, based on stellar parallaxes then just detected, his geometry could be physically meaningful only in gargantuan triangles spanning interstellar space.

In fact, non-Euclidean geometries apply to the cosmos more locally than Lobachevsky imagined. In 1916 Albert Einstein (1879–1955) published "The Foundation of the General Theory of Relativity," which replaced Newton's description of gravitation as a force that attracts distant masses to each other through Euclidean space with a principle of least effort, or shortest (temporal) path, for motion along the geodesics of a curved space. Einstein not only explained how gravitating bodies give this surface its properties—that is, mass determines how the differential distances, or curvatures, in Riemann's geometry differ from those in Euclidean space—but also successfully predicted the deflection of light, which has no mass, in the vicinity of a star or other massive body. This was an extravagant piece of geometrizing—the replacement of gravitational force by the curvature of a surface. But it was not all. In relativity theory time is considered to be a dimension along with the three dimensions of space. On the closed four-dimensional world thus formed, the history of the universe stands revealed as describable by motion within a vast congeries of geodesics in a non-Euclidean universe. (J.L.He.)

Einstein's space-time model

BRANCHES OF GEOMETRY

Euclidean geometry

Euclidean geometry is the study of plane and solid figures on the basis of axioms and theorems employed by the Greek mathematician Euclid (*c.* 300 BC). In its rough outline, Euclidean geometry is the plane and solid geometry commonly taught in secondary schools. Indeed, until the second half of the 19th century, when non-Euclidean geometries attracted the attention of mathematicians, geometry meant Euclidean geometry. It is the most typical expression of general mathematical thinking. Rather than the memorization of simple algorithms to solve equations by rote, it demands true insight into the subject, clever ideas for applying theorems in special situations, an ability to generalize from known facts, and an insistence on the importance of proof. In Euclid's great work, the *Elements*, the only tools employed for geometrical constructions were the ruler and compass—a restriction retained in elementary Euclidean geometry to this day.

In its rigorous deductive organization, the *Elements* remained the very model of scientific exposition until the end of the 19th century, when the German mathematician David Hilbert wrote his famous *Foundations of Geometry* (1899). The modern version of Euclidean geometry is the theory of Euclidean (coordinate) spaces of multiple dimensions, where distance is measured by a suitable generalization of the Pythagorean theorem.

FUNDAMENTALS

Euclid realized that a rigorous development of geometry must start with the foundations. Hence, he began the *Elements* with some undefined terms, such as "a point is that which has no part" and "a line is a length without breadth." Proceeding from these terms, he defined further ideas such as angles, circles, triangles, and various other polygons and figures. For example, an angle was defined as the inclination of two straight lines, and a circle was a plane figure consisting of all points that have a fixed distance (radius) from a given centre.

Euclid's five axioms

As a basis for further logical deductions, Euclid proposed five common notions, such as "things equal to the same thing are equal," and five unprovable but intuitive principles known variously as postulates or axioms. Stated in modern terms, the axioms are as follows:

1. Given two points, there is a straight line that joins them.
2. A straight line segment can be prolonged indefinitely.
3. A circle can be constructed when a point for its centre and a distance for its radius are given.
4. All right angles are equal.
5. If a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, will meet on that side on which the angles are less than the two right angles.

Hilbert refined axioms (1) and (5) as follows:

1. For any two different points, (a) there exists a line containing these two points, and (b) this line is unique.

5. For any line L and point p not on L, (a) there exists a line through p not meeting L, and (b) this line is unique.

The fifth axiom became known as the "parallel postulate," since it provided a basis for the uniqueness of parallel lines. (It also attracted great interest because it seemed less intuitive or self-evident than the others. In the 19th century, Carl Friedrich Gauss, János Bolyai, and Nikolay Lobachevsky all began to experiment with this postulate, eventually arriving at new, non-Euclidean, geometries.) All five axioms provided the basis for numerous provable statements, or theorems, on which Euclid built his geometry. The rest of this section briefly explains the most important theorems of Euclidean plane and solid geometry.

PLANE GEOMETRY

Congruence of triangles. Two triangles are said to be congruent if one can be exactly superimposed on the other by a rigid motion, and the congruence theorems specify the conditions under which this can occur. The first theorem illustrated in Figure 28 is the side-angle-side (SAS) theorem: If two sides and the included angle of one triangle are equal to two sides and the included angle of another triangle, the triangles are congruent. Following this, there are corresponding angle-side-angle (ASA) and side-side-side (SSS) theorems.

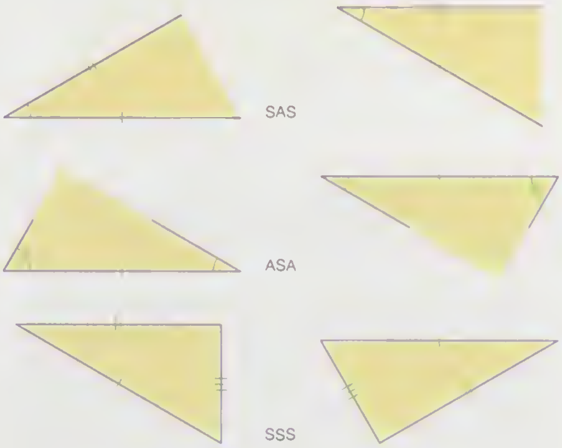


Figure 28: Congruent triangle theorems.

The first very useful theorem derived from the axioms is the basic symmetry property of isosceles triangles, *i.e.*, that two sides of a triangle are equal if and only if the angles opposite them are equal. Euclid's proof of this theorem was once called *Pons Asinorum* ("Bridge of Asses"), supposedly because mediocre students could not proceed across it to the farther reaches of geometry (see Figure 10 in *History of geometry: Idealization and proof,* above). The Bridge of Asses opens the way to various theorems on the congruence of triangles.

The parallel postulate is fundamental for the proof of the theorem that the sum of the angles of a triangle is always 180 degrees. A simple proof of this theorem, attributed to the Pythagoreans, is shown in Figure 29.

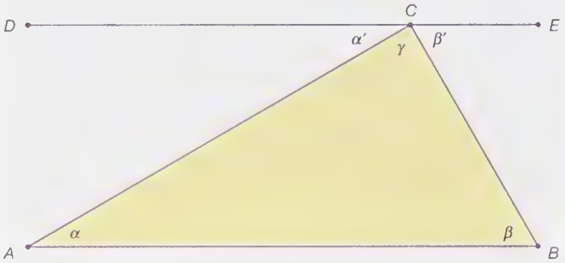


Figure 29: *The sum of the angles in a triangle is 180 degrees.* According to an ancient theorem, a transversal through two parallel lines (*DE* and *AB* in the figure) forms several equal angles, such as the alternating angles α/α' and β/β', labeled in the figure. By definition, the three angles α', γ, and β' on the line *DE* must sum to 180 degrees. Since $\alpha = \alpha'$ and $\beta = \beta'$, the sum of the angles in the triangle (α, β, and γ) is also 180 degrees.

Similarity of triangles. As indicated above, congruent figures have the same shape and size. Similar figures, on the other hand, have the same shape but may differ in size. Shape is intimately related to the notion of proportion, as ancient Egyptian artisans observed long ago. Segments of lengths a, b, c, and d are said to be proportional if $a{:}b = c{:}d$ (read, a is to b as c is to d; in older notation $a{:}b{::}c{:}d$). The fundamental theorem of similarity states that a line segment splits two sides of a triangle into proportional segments if and only if the segment is parallel to the triangle's third side (see Figure 30).

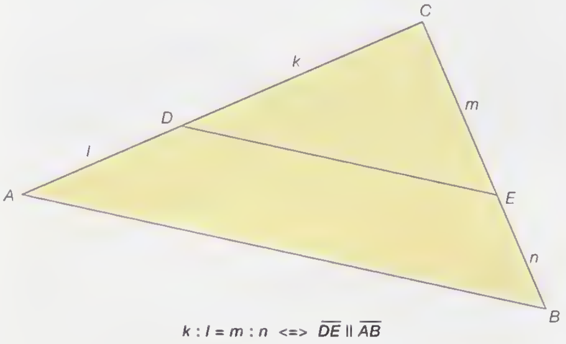
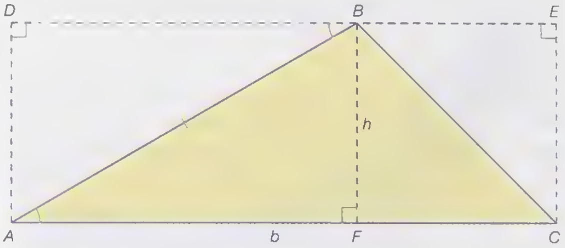


Figure 30: *The fundamental theorem of similarity.* The formula in the figure reads *k* is to *l* as *m* is to *n* if and only if line *DE* is parallel to the line *AB*. This theorem then enables one to show that the small and large triangles are similar.

The similarity theorem may be reformulated as the AAA (angle-angle-angle) similarity theorem: two triangles have their corresponding angles equal if and only if their corresponding sides are proportional. Two similar triangles are related by a scaling (or similarity) factor s: if the first triangle has sides a, b, and c, then the second one will have sides sa, sb, and sc. In addition to the ubiquitous use of scaling factors on construction plans and geographic maps, similarity is fundamental to trigonometry.

Areas. Just as a segment can be measured by comparing it with a unit segment, the area of a polygon or other plane figure can be measured by comparing it with a unit square. The common formulas for calculating areas reduce this kind of measurement to the measurement of certain suitable lengths. The simplcst case is a rectangle with sides a and b, which has area ab. By putting a triangle into an appropriate rectangle (see Figure 31), one can show that the area of the triangle is half the product of the length of one of its bases and its corresponding height—$bh/2$. One can then compute the area of a general polygon by dissecting it into triangular regions. If a triangle (or more general figure) has area A, a similar triangle (or figure) with a scaling factor of s will have an area of s^2A.



The right triangle $\triangle AFB$ is $\frac{1}{2}$ of the rectangle □$ADBF$.
Similarly, $\triangle BFC$ is $\frac{1}{2}$ of □$BECF$.
Thus, the area of $\triangle ABC = \frac{1}{2}$ area of □$ADEC = \frac{1}{2}\ AC \cdot BF = \frac{1}{2}$ base · height. ■

Figure 31: Area of a triangle.

Pythagorean theorem. For a triangle ΔABC the Pythagorean theorem has two parts: (1) if $\angle ACB$ is a right angle, then $a^2 + b^2 = c^2$; (2) if $a^2 + b^2 = c^2$, then $\angle ACB$ is a right angle. For an arbitrary triangle, the Pythagorean theorem is generalized to the law of cosines: $a^2 + b^2 = c^2 - 2ab$ cos ($\angle ACB$). When $\angle ACB$ is 90 degrees, this reduces to the Pythagorean theorem because cos (90°) = 0.

Law of cosines

Since Euclid, a host of professional and amateur mathematicians have found more than 300 distinct proofs of the Pythagorean theorem. Despite its antiquity, it remains one of the most important theorems in mathematics. It enables one to calculate distances or, more importantly, to define distances in situations far more general than elementary geometry. For example, it has been generalized to multidimensional vector spaces.

Circles. A chord AB is a segment in the interior of a circle connecting two points (A and B) on the circumference. When a chord passes through the circle's centre, it is a diameter, d. The circumference of a circle is given by πd, or $2\pi r$ where r is the radius of the circle; the area of a circle is πr^2. In each case, π is the same constant (3.14159 . . .). The Greek mathematician Archimedes (*c.* 285–212/211 BC) used the method of exhaustion to obtain upper and lower bounds for π by circumscribing and inscribing regular polygons about a circle (see Figure 9).

A semicircle has its end points on a diameter of a circle. Thales (fl. 6th century BC) is generally credited with proving that any angle inscribed in a semicircle is a right angle; that is, for any point C on the semicircle with diameter AB, $\angle ACB$ will always be 90 degrees. Another important theorem states that for any chord AB in a circle, the angle subtended by any point on the same semiarc of the circle will be invariant (see Figure 32). Slightly modified, this means that in a circle, equal chords determine equal angles, and vice versa.

Summarizing the above material, the five most important theorems of plane Euclidean geometry are: the sum of the angles in a triangle is 180 degrees, the Bridge of Asses, the fundamental theorem of similarity, the Pythagorean theorem, and the invariance of angles subtended by a chord in

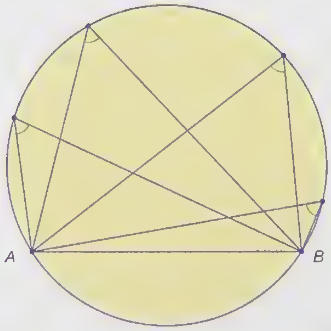


Figure 32: Invariance of subtended angles on a chord.

a circle. Most of the more advanced theorems of plane Euclidean geometry are proved with the help of these theorems.

Regular polygons. A polygon is called regular if it has equal sides and angles. Thus, a regular triangle is an equilateral triangle, and a regular quadrilateral is a square. A general problem since antiquity has been the problem of constructing a regular n-gon, for different n, with only ruler and compass. For example, Euclid constructed a regular pentagon by applying the above-mentioned five important theorems in an ingenious combination.

Techniques, such as bisecting the angles of known constructions, exist for constructing regular n-gons for many values, but none is known for the general case. In 1797, following centuries without any progress, Gauss surprised the mathematical community by discovering a construction for the 17-gon. More generally, Gauss was able to show that for a prime number p, the regular p-gon is constructible if and only if p is a "Fermat prime": $p = F(k) = 2^{2^k} + 1$. Because it is not known in general which $F(k)$ are prime, the construction problem for regular n-gons is still open.

Fermat primes

Three other unsolved construction problems from antiquity were finally settled in the 19th century by applying tools not available to the Greeks. Comparatively simple algebraic methods showed that it is not possible to trisect an angle with ruler and compass, or to construct a cube with a volume double that of a given cube. To show that it is not possible to square a circle, *i.e.*, to construct a square equal in area to a given circle by the same means, however, demanded deeper insights into the nature of the number π.

Conic sections and geometric art. The most advanced part of plane Euclidean geometry is the theory of the conic sections (the ellipse, parabola, and hyperbola; see Figure 12). Much as the *Elements* displaced all other introductions to geometry, the *Conics* of Apollonius of Perga (*c.* 240–190 BC), known by his contemporaries as "The Great Geometer," was for many centuries the definitive treatise on the subject.

Islāmic artists explored ways of using geometric figures for decoration. For example, the decorations of the Alhambra of Granada, Spain, demonstrate an understanding of all 17 of the different "Wallpaper groups" that can be used to tile the plane. In the 20th century, internationally renowned artists such as Josef Albers, Max Bill, and Sol Le Witt were inspired by motifs from Euclidean geometry.

SOLID GEOMETRY

The most important difference between plane and solid Euclidean geometry is that human beings can look at the plane "from above," whereas three-dimensional space cannot be looked at "from outside." Consequently, intuitive insights are more difficult to obtain for solid geometry than for plane geometry.

Some concepts, such as proportions and angles, remain unchanged from plane to solid geometry. For other familiar concepts, there exist analogies—most noticeably, volume for area and three-dimensional shapes for two-dimensional shapes (sphere for circle, tetrahedron for triangle, box for rectangle). However, the theory of tetrahedra is not nearly as rich as it is for triangles. Active research in higher-dimensional Euclidean geometry includes convexity and sphere packings and their applications in cryptology and crystallography.

Volume. As is explained above, in plane geometry the area of any polygon can be calculated by dissecting it into triangles. A similar procedure is not possible for solids. In 1901 the German mathematician Max Dehn showed that there exist a cube and a tetrahedron of equal volume that cannot be dissected and rearranged into each other. This means that calculus must be used to calculate volumes for even many simple solids like pyramids.

Regular solids. Regular polyhedra are the solid analogies to regular polygons in the plane. Regular polygons are defined as having equal (congruent) sides and angles. In analogy, a solid is called regular if its faces are congruent regular polygons and its polyhedral angles (angles at which the faces meet) are congruent. This concept has been generalized to higher-dimensional (coordinate) Euclidean spaces.

Whereas in the plane there exist (in theory) infinitely many regular polygons, in three-dimensional space there exist exactly five regular polyhedra. These are known as the Platonic solids: the tetrahedron, or pyramid, with 4 triangular faces; the cube, with 6 square faces; the octahedron, with 8 equilateral triangular faces; the dodecahedron, with 12 pentagonal faces; and the icosahedron, with 20 equilateral triangular faces. (See Figure 14.)

Platonic solids

In four-dimensional space there exist exactly six regular polytopes, five of them generalizations from three-dimensional space. In any space of more than four dimensions there exist exactly three regular polytopes, the generalizations of the tetrahedron, the cube, and the octahedron.

(B.Ar.)

Projective geometry

Projective geometry is the branch of mathematics that deals with the relationships between geometric figures and the images, or mappings, that result from projecting them onto another surface. Common examples of projections are the shadows cast by opaque objects and motion pictures displayed on a screen.

Projective geometry has its origins in the early Italian Renaissance, particularly in the architectural drawings of Filippo Brunelleschi (1377–1446) and Leon Battista Alberti (1404–1472), who invented the method of perspective drawing. By this method, as shown in Figure 33, the eye of the painter is connected to points on the landscape (the horizontal reality plane, *RP*) by so-called sight lines. The intersection of these sight lines with the vertical picture plane (*PP*) generates the drawing. Thus, the reality plane is projected onto the picture plane, hence the name projective geometry.

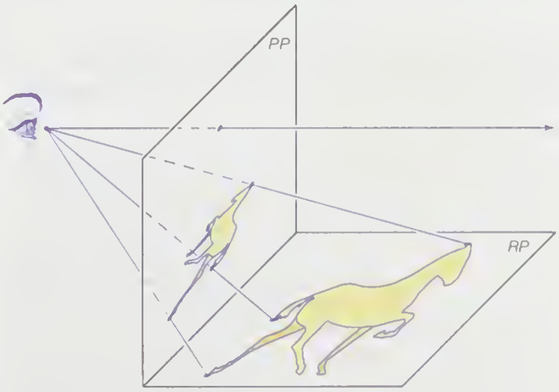


Figure 33: Projective drawing.

Although some isolated properties concerning projections were known in antiquity, particularly in the study of optics, it was not until the 17th century that mathematicians returned to the subject. The French mathematicians Girard Desargues (1591–1661) and Blaise Pascal (1623–62) took the first significant steps by examining what properties of figures were preserved (or invariant) under perspective mappings. The subject's real importance, however, became clear only after 1800 in the works of several other French mathematicians, notably Jean-Victor Poncelet (1788–1867). In general, by ignoring geometric measurements such as distances and angles, projective geometry enables a clearer understanding of some more generic properties of geometric objects. Such insights have since been incorporated in many more advanced areas of mathematics.

PARALLEL LINES AND THE PROJECTION OF INFINITY

A theorem from Euclid's *Elements* (*c.* 300 BC) states that if a line is drawn through a triangle such that it is parallel to one side (see Figure 30), then the line will divide the other two sides proportionately—that is, the ratio of segments on each side will be equal. This is known as the proportional segments theorem, or the fundamental theorem of similar-

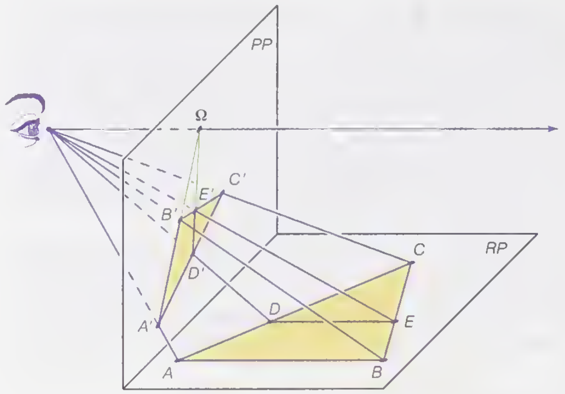


Figure 34: Similarity under projection.

ity, and for triangle ABC, shown in Figure 34, with line segment DE parallel to side AB, the theorem corresponds to the mathematical expression $CD/DA = CE/EB$.

Single point at infinity

Now consider the effect produced by projecting these line segments onto another plane. The first thing to note is that the projected line segments $A'B'$ and $D'E'$ are not parallel—*i.e.*, angles are not preserved. From the point of view of the projection, the parallel lines AB and DE appear to converge at the horizon, or at infinity, whose projection in the picture plane is labeled Ω. (It was Desargues who first introduced a single point at infinity to represent the projected intersection of parallel lines. Furthermore, he collected all the points along the horizon in one line at infinity.) With the introduction of Ω, the projected figure corresponds to a theorem discovered by Menelaus of Alexandria in the 1st century AD:

$$\frac{C'D'}{D'A'} = \frac{C'E'}{E'B'} \cdot \frac{\Omega B'}{\Omega A'}$$

Since the factor $\Omega B'/\Omega A'$ corrects for the projective distortion in lengths, Menelaus's theorem can be seen as a projective variant of the proportional segments theorem.

PROJECTIVE INVARIANTS

With Desargues's provision of infinitely distant points for parallels, the reality plane and the projective plane are essentially interchangeable—that is, ignoring distances and directions (angles), which are not preserved in the projection. Other properties are preserved, however. For instance, two different points have a unique connecting line, and two different lines have a unique point of intersection. Although almost nothing else seems to be invariant under projective mappings, one should note that lines are mapped onto lines. This means that if three points are collinear (share a common line), then the same will be true for their projections. Thus, collinearity is another invariant property. Similarly, if three lines meet in a common point, so will their projections.

The following theorem is of fundamental importance for projective geometry. In its first variant, by Pappus of Alexandria (*c.* AD 320) as shown in Figure 35, it only uses collinearity:

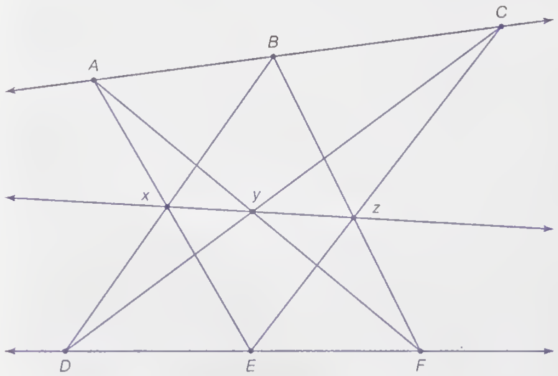


Figure 35: Pappus's theorem.

Let the distinct points A, B, C and D, E, F be on two different lines. Then the three intersection points—x of AE and BD, y of AF and CD, and z of BF and CE—are collinear.

The second variant, by Pascal, as shown in Figure 36, uses certain properties of circles:

If the distinct points A, B, C, D, E, and F are on one circle, then the three intersection points x, y, and z (defined as above) are collinear.

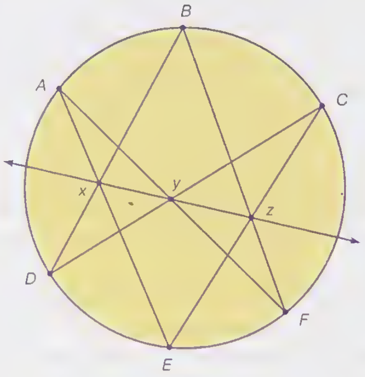


Figure 36: Pascal's projective theorem.

There is one more important invariant under projective mappings, known as the cross ratio (see Figure 37). Given four distinct collinear points A, B, C, and D, the cross ratio is defined as $\mathrm{CRat}(A, B, C, D) = AC/BC \cdot BD/AD$. It may also be written as the quotient of two ratios:

$$\mathrm{CRat}(A, B, C, D) = AC/BC : AD/BD$$

The latter formulation reveals the cross ratio as a ratio of ratios of distances. And while neither distance nor the ratio of distance is preserved under projection, Pappus proved the startling fact that the cross ratio was invariant—that is, $\mathrm{CRat}(A, B, C, D) = \mathrm{CRat}(A', B', C', D')$. However, this result remained a mere curiosity until its real significance became gradually clear in the 19th century as mappings became more and more important for transforming problems from one mathematical domain to another.

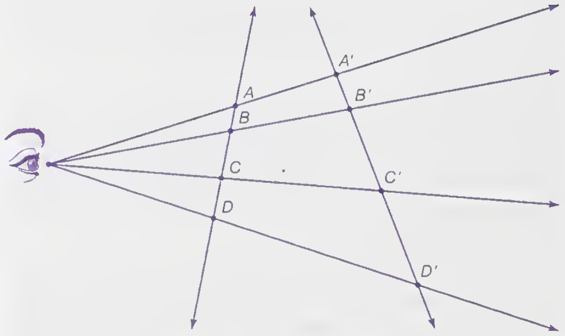


Figure 37: Cross ratio.

PROJECTIVE CONIC SECTIONS

Conic sections can be regarded as plane sections of a right circular cone. By regarding a plane perpendicular to the cone's axis as the reality plane (RP), a "cutting" plane as the picture plane (PP), and the cone's apex as the projective "eye," each conic section can be seen to correspond to a projective image of a circle (see Figure 38). Depending on the orientation of the cutting plane, the image of the circle will be a circle, an ellipse, a parabola, or a hyperbola.

A plane Ω passing through the apex and parallel to PP defines the line at infinity in the projective plane PP. The situation of Ω relative to RP determines the conic section in PP: If Ω intersects RP outside the base circle (the circle formed by the intersection of the cone and RP), the image of the circle will be an ellipse (as shown in Figure 38). If Ω is tangent to the base circle (in effect, tangent to the cone), the image will be a parabola. If Ω intersects the base circle (thus, cutting the circle in two), a hyperbola will result.

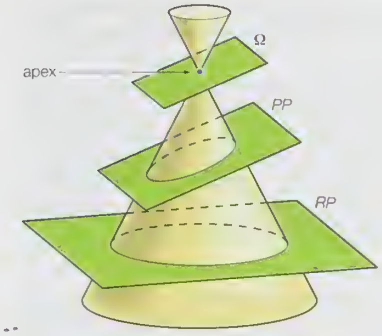


Figure 38: Conic sections as projections.

Pascal's theorem, quoted above, also follows easily for any conic section from its special case for the circle. Start by selecting six points on a conic section and project them back onto the base circle. As given earlier, the three relevant intersection points for six points on the circle will be collinear. Now project all nine points back to the conic section. Since collinear points (the three intersection points from the circle) are mapped onto collinear points, the theorem holds for any conic section. In this way the projective point of view unites the three different types of conics.

Similarly, more complicated curves and surfaces in higher-dimensional spaces can be unified through projections. For example, Isaac Newton (1643–1727) showed that all plane curves defined by polynomials in x and y of degree 3 (the highest power of the variables is 3) can be obtained as projective images of just five types of polynomials.

(B.Ar.)

Differential geometry

Differential geometry is the branch of mathematics that studies the geometry of curves, surfaces, and manifolds (the higher-dimensional analogs of surfaces). The discipline owes its name to its use of ideas and techniques from differential calculus, though the modern subject often uses algebraic and purely geometric techniques instead. Although basic definitions, notations, and analytic descriptions vary widely, the following geometric questions prevail: How does one measure the curvature of a curve within a surface (intrinsic) versus within the encompassing space (extrinsic)? How can the curvature of a surface be measured? What is the shortest path within a surface between two points on the surface? How is the shortest path on a surface related to the concept of a straight line?

While curves had been studied since antiquity, the discovery of calculus in the 17th century opened up the study of more complicated plane curves—such as those produced by the French mathematician René Descartes (1596–1650) with his "compass." In particular, integral calculus led to general solutions of the ancient problems of finding the arc length of plane curves and the area of plane figures. This in turn opened the stage to the investigation of curves and surfaces in space—an investigation that was the start of differential geometry.

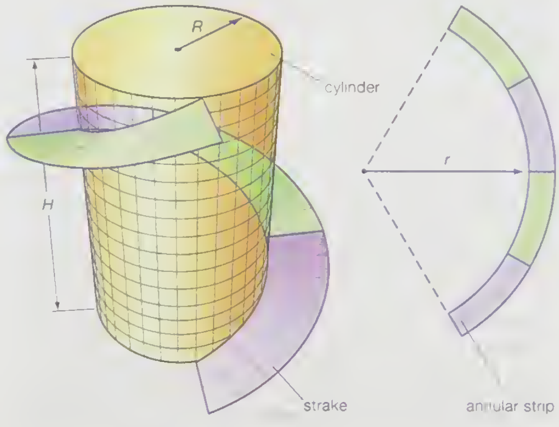


Figure 39: Forming a strake.

Some of the fundamental ideas of differential geometry can be illustrated by the strake, a spiraling strip often designed by engineers to give structural support to large metal cylinders such as smokestacks. A strake can be formed by cutting an annular strip (the region between two concentric circles) from a flat sheet of steel and then bending it into a helix that spirals around the cylinder, as illustrated in Figure 39. What should the radius r of the annulus be to produce the best fit? Differential geometry supplies the solution to this problem by defining a precise measurement for the curvature of a curve; then r can be adjusted until the curvature of the inside edge of the annulus matches the curvature of the helix.

An important question remains: Can the annular strip be bent, without stretching, so that it forms a strake around the cylinder? In particular, this means that distances measured along the surface (intrinsic) are unchanged. Two surfaces are said to be isometric if one can be bent (or transformed) into the other without changing intrinsic distances. (For example, because a sheet of paper can be rolled into a tube without stretching, the sheet and tube are "locally" isometric—only locally because new, and possibly shorter, routes are created by connecting the two edges of the paper.) Thus, the second question becomes: Are the annular strip and the strake isometric? To answer this and similar questions, differential geometry developed the notion of the curvature of a surface.

Isometric surfaces

CURVATURE OF CURVES

Although mathematicians from antiquity had described some curves as curving more than others and straight lines as not curving at all, it was the German mathematician Gottfried Leibniz who, in 1686, first defined the curvature of a curve at each point in terms of the circle that best approximates the curve at that point. As shown in Figure 40, Leibniz named his approximating circle the osculating circle, from the Latin *osculare* ("to kiss"). He then defined the curvature of the curve (and the circle) as $1/r$, where r is the radius of the osculating circle. As a curve becomes straighter, a circle with a larger radius must be used to approximate it, and so the resulting curvature decreases. In the limit, a straight line is said to be equivalent to a circle of infinite radius and its curvature defined as zero everywhere. The only curves in ordinary Euclidean space with constant curvature are straight lines, circles, and helices. In practice, curvature is found with a formula that gives the rate of change, or derivative, of the tangent to the curve as one moves along the curve. This formula was discovered by Isaac Newton and Leibniz for plane curves in the 17th century and by the Swiss mathematician Leonhard Euler for curves in space in the 18th century. (Note that the derivative of the tangent to the curve is not the same as the second derivative studied in calculus, which is the rate of change of the tangent to the curve as one moves along the x-axis.)

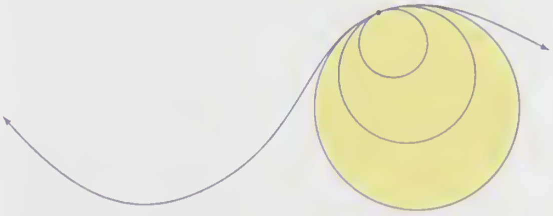


Figure 40: *Osculating circles.*
Note that while all of the circles in the figure are tangent to the curve at the same point, only the largest circle is a true osculating circle.

With these definitions in place, it is now possible to compute the ideal inner radius r of the annular strip that goes into making the strake shown in Figure 39. The annular strip's inner curvature $1/r$ must equal the curvature of the helix on the cylinder. If R is the radius of the cylinder and H is the height of one turn of the helix, then the curvature of the helix is $4\pi^2R/[H^2 + (2\pi R)^2]$. For example, if $R = 1$ metre and $H = 10$ metres, then $r = 3.533$ metres.

CURVATURE OF SURFACES

To measure the curvature of a surface at a point, Euler, in 1760, looked at cross sections of the surface made by planes that contain the line perpendicular (or "normal") to the surface at the point (see Figure 41). Euler called the curvatures of these cross sections the normal curvatures of the surface at the point. For example, on a right cylinder of radius r, the vertical cross sections are straight lines and thus have zero curvature; the horizontal cross sections are circles, which have curvature $1/r$. The normal curvatures at a point on a surface are generally different in different directions. The maximum and minimum normal curvatures at a point on a surface are called the principal (normal) curvatures, and the directions in which these normal curvatures occur are called the principal directions. Euler proved that for most surfaces where the normal curvatures are not constant (for example, the cylinder), these principal directions are perpendicular to each other. (Note that on a sphere all the normal curvatures are the same and thus all are principal curvatures.) These principal normal curvatures are a measure of how "curvy" the surface is.

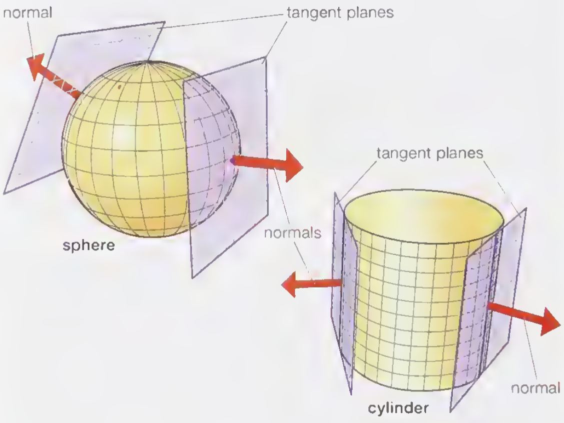


Figure 41: Normals to a surface.

Gaussian curvature

The theory of surfaces and principal normal curvatures was extensively developed by French geometers led by Gaspard Monge (1746–1818). It was in an 1827 paper, however, that the German mathematician Carl Freidrich Gauss made the big breakthrough that allowed differential geometry to answer the question raised above of whether the annular strip is isometric to the strake. The Gaussian curvature of a surface at a point is defined as the product of the two principal normal curvatures; it is said to be positive if the principal normal curvatures curve in the same direction and negative if they curve in opposite directions. Normal curvatures for a plane surface are all zero, and thus the Gaussian curvature of a plane is zero. For a cylinder of radius r, the minimum normal curvature is zero (along the vertical straight lines), and the maximum is $1/r$ (along the horizontal circles). Thus, the Gaussian curvature of a cylinder is also zero.

If the cylinder is cut along one of the vertical straight lines, the resulting surface can be flattened (without stretching) onto a rectangle. In differential geometry, it is said that the plane and cylinder are locally isometric. These are special cases of two important theorems: Gauss's "Remarkable Theorem" (1827) states that if two smooth surfaces are isometric, then the two surfaces have the same Gaussian curvature at corresponding points. (Although defined extrinsically, Gaussian curvature is an intrinsic notion.)

Minding's theorem (1839) states that two smooth surfaces ("cornerless") with the same constant Gaussian curvature are locally isometric.

As corollaries to these theorems:

> A surface with constant positive Gaussian curvature c has locally the same intrinsic geometry as a sphere of radius $\sqrt{1/c}$. (This is because a sphere of radius r has Gaussian curvature $1/r^2$.)
>
> A surface with constant zero Gaussian curvature has locally the same intrinsic geometry as a plane. (Such surfaces are called developable.)
>
> A surface with constant negative Gaussian curvature c has locally the same intrinsic geometry as a hyperbolic plane.

The Gaussian curvature of an annular strip (being in the plane) is constantly zero. So to answer whether or not the annular strip is isometric to the strake, one needs only to check whether a strake has constant zero Gaussian curvature. The Gaussian curvature of a strake is actually negative, hence the annular strip must be stretched—although this can be minimized by narrowing the shapes.

SHORTEST PATHS ON A SURFACE

From an outside, or extrinsic, perspective, no curve on a sphere is straight. Nevertheless, the great circles are intrinsically straight—an ant crawling along a great circle does not turn or curve with respect to the surface. In about 1830 the Estonian mathematician Ferdinand Minding defined a curve on a surface to be a geodesic if it is intrinsically straight; that is, if there is no identifiable curvature from within the surface. A major task of differential geometry is to determine the geodesics on a surface. The great circles are the geodesics on a sphere.

A great circle arc that is longer than a half circle is intrinsically straight on the sphere, but it is not the shortest distance between its endpoints. On the other hand, the shortest path in a surface is not always straight, as shown in Figure 42. An important theorem is:

> On a surface which is complete (every geodesic can be extended indefinitely) and smooth, every shortest curve is intrinsically straight and every intrinsically straight curve is the shortest curve between nearby points.

(D.W.H.)

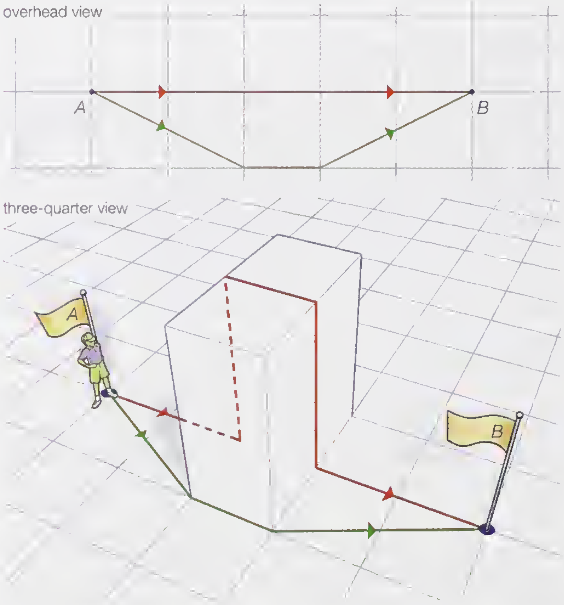


Figure 42: *Shortest paths are not always straight.* Although the red path (which appears straight from overhead) is intrinsically straight, the green path (around the elevation) is a shorter path.

Analytic geometry

Analytic geometry, also called coordinate geometry, is the mathematical subject in which algebraic symbolism and methods are used to represent and solve problems in geometry. The importance of analytic geometry is that it establishes a correspondence between geometric curves and algebraic equations. This correspondence makes it possible to reformulate problems in geometry as equivalent problems in algebra, and vice versa; the methods of either subject can then be used to solve problems in the other. For example, computers create animations for display in games and films by manipulating algebraic equations.

Analytic geometry was once called Cartesian geometry, after René Descartes, who described the basic approach in an appendix to his *Discourse on Method* (1937).

ELEMENTARY ANALYTIC GEOMETRY

Coordinate systems and algebraic symbolism. Apollonius of Perga (*c.* 262–190 BC), known by his contemporaries as "The Great Geometer," foreshadowed the development of analytic geometry by more than 1,800 years with his book *Conics*. He defined a conic as the intersection of a cone and a plane (see Figure 12). Using Euclid's results on similar triangles and on secants of circles, he found a relation satisfied by the distances from any point P of a conic to two perpendicular lines, the major axis of the conic and the tangent at an endpoint of the axis. These distances correspond to coordinates of P, and the relation between these coordinates corresponds to a quadratic equation of the conic. Apollonius used this relation to deduce fundamental properties of conics.

Viète and algebraic symbolism

Further development of coordinate systems (see Figure 43) in mathematics emerged only after algebra had matured under Islāmic and Indian mathematicians. At the end of the 16th century, the French mathematician François Viète introduced the first systematic algebraic notation, using letters to represent known and unknown numerical quantities, and he developed powerful general methods for working with algebraic expressions and solving algebraic equations. With the power of algebraic notation, mathematicians were no longer completely dependent upon geometric figures and geometric intuition to solve problems. The more daring began to leave behind the standard geometric way of thinking in which linear (1st power) variables corresponded to lengths, squares (2nd power) to areas, and cubics (3rd power) to volumes, with higher powers lacking "physical" interpretation. Two Frenchmen, the mathematician-philosopher René Descartes and the lawyer-mathematician Pierre de Fermat, were among the first to take this daring step.

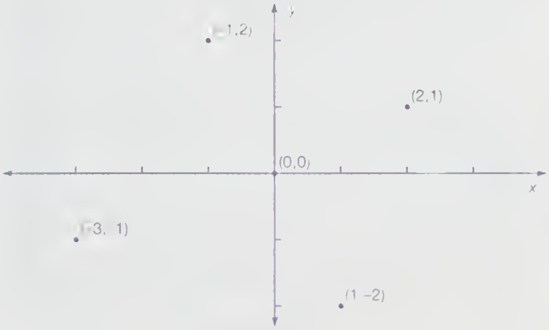


Figure 43: *Cartesian coordinates.*
Several points are labeled in a two-dimensional graph, known as the Cartesian plane. Note that each point has two coordinates, the first number (x value) indicates its distance from the y-axis—positive values to the right and negative values to the left—and the second number (y value) gives its distance from the x-axis—positive values upward and negative values downward.

Descartes and Fermat independently founded analytic geometry in the 1630s by adapting Viète's algebra to the study of geometric loci. They moved decisively beyond Viète by using letters to represent distances that are variable instead of fixed. Descartes used equations to study curves defined geometrically, and he stressed the need to consider general algebraic curves—graphs of polynomial equations in x and y of all degrees. He demonstrated his method on a classical problem: finding all points P such that the product of the distances from P to certain lines equals the product of the distances to other lines.

Fermat emphasized that any relation between x and y coordinates determines a curve (see Figure 44). Using this idea, he recast Apollonius's arguments in algebraic terms and restored lost work. Fermat indicated that any quadratic equation in x and y can be put into the standard form of one of the conic sections.

Fermat did not publish his work, and Descartes deliberately made his hard to read in order to discourage "dabblers." Their ideas gained general acceptance only through the efforts of other mathematicians in the latter half of the 17th century. In particular, the Dutch mathematician Frans van Schooten translated Descartes's writings from French to Latin. He added vital explanatory material, as did the French lawyer Florimond de Beaune and the Dutch mathematician Johan de Witt. In England, the mathematician John Wallis popularized analytic geometry, using equations to define conics and derive their properties. He used negative coordinates freely, although it was Isaac Newton who unequivocally used two (oblique) axes to divide the plane into four quadrants, as shown in Figure 43.

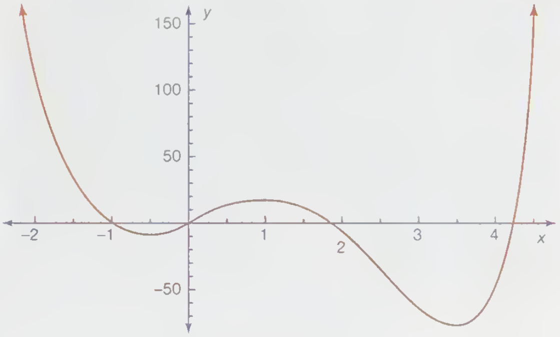


Figure 44: *Polynomial graph.*
The figure shows part of the graph of the polynomial equation $y = 3x^4 - 16x^3 + 6x^2 + 24x + 1$. Note that the same scale need not be used for the x- and y-axis.

Analytic geometry and the development of calculus. Analytic geometry had its greatest impact on mathematics via calculus. Without access to the power of analytic geometry, classical Greek mathematicians such as Archimedes (*c.* 285–212/211 BC) solved special cases of the basic problems of calculus: finding tangents and extreme points (differential calculus) and arc lengths, areas, and volumes (integral calculus). Renaissance mathematicians were led back to these problems by the needs of astronomy, optics, navigation, warfare, and commerce. They naturally sought to use the power of algebra to define and analyze a growing range of curves.

Descartes and differential calculus

Fermat developed an algebraic algorithm for finding the tangent to an algebraic curve at a point by finding a line that has a double intersection with the curve at the point—in essence, inventing differential calculus. Descartes introduced a similar but more complicated algorithm using a circle. Fermat computed areas under the curves $y = ax^k$ for all rational numbers $k \neq -1$ by summing areas of inscribed and circumscribed rectangles. For the rest of the 17th century, the groundwork for calculus was continued by many mathematicians, including the Frenchman Gilles Personne de Roberval, the Italian Bonaventura Cavalieri, and the Britons James Gregory, John Wallis, and Isaac Barrow.

Newton and the German Gottfried Leibniz revolutionized mathematics at the end of the 17th century by independently demonstrating the power of calculus. Both men used coordinates to develop notations that expressed the ideas of calculus in full generality and led naturally to differentiation rules and the fundamental theorem of calculus (connecting differential and integral calculus).

Newton demonstrated the importance of analytic methods in geometry, apart from their role in calculus, when he asserted that any cubic—algebraic curve of degree three—has one of four standard equations

$$\begin{aligned} xy^2 + ey &= ax^3 + bx^2 + cx + d, \\ xy &= ax^3 + bx^2 + cx + d, \\ y^2 &= ax^3 + bx^2 + cx + d, \\ y &= ax^3 + bx^2 + cx + d \end{aligned}$$

for suitable coordinate axes. The Scottish mathematician James Stirling proved this assertion in 1717, possibly with Newton's aid. Newton divided cubics into 72 species, a total later corrected to 78.

Newton also showed how to express an algebraic curve near the origin in terms of fractional power series $y = a_1x^{1/k} + a_2x^{2/k} + \dots$ for a positive integer k. Mathematicians have since used this technique to study algebraic curves of all degrees.

ANALYTIC GEOMETRY OF THREE OR MORE DIMENSIONS

Although both Descartes and Fermat suggested using three coordinates to study curves and surfaces in space, three-dimensional analytic geometry developed slowly until about 1730, when the Swiss mathematicians Leonhard Euler and Jakob Hermann and the French mathematician Alexis Clairaut produced general equations for cylinders, cones, and surfaces of revolution. For example, Euler and Hermann showed that the equation $f(z) = x^2 + y^2$ gives the surface that is produced by revolving the curve $f(z) = x^2$ about the z-axis (see Figure 45, which gives the elliptic paraboloid $z = x^2 + y^2$).

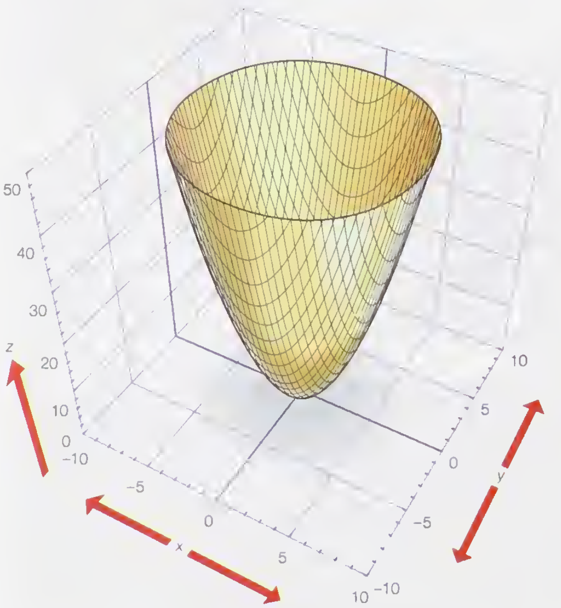


Figure 45: *Elliptic paraboloid.*
The figure shows part of the elliptic paraboloid $z = x^2 + y^2$, which can be generated by rotating the parabola $z = x^2$ (or $z = y^2$) about the z-axis. Note that cross sections of the surface parallel to the xy plane, as shown by the cutoff at the top of the figure, are ellipses.

Newton made the remarkable claim that all plane cubics arise from those in his third standard form by projection between planes. This was proved independently in 1731 by Clairaut and the French mathematician François Nicole. Clairaut obtained all the cubics in Newton's four standard forms as sections of the cubical cone

$$zy^2 = ax^3 + bx^2z + cxz^2 + dz^3$$

consisting of the lines in space that join the origin (0, 0, 0) to the points on the third standard cubic in the plane $z = 1$.

In 1748 Euler used equations for rotations and translations in space to transform the general quadric surface

$$ax^2 + by^2 + cz^2 + dxy + exz + fyz + gx + hy + iz + j = 0$$

so that its principal axes coincide with the coordinate axes. Euler and the French mathematicians Joseph-Louis Lagrange and Gaspard Monge made analytic geometry independent of synthetic (nonanalytic) geometry.

Vector analysis. In Euclidean space of any dimension, vectors—directed line segments—can be specified by coordinates. An n-tuple $(a_1, \ldots, a_n)$ represents the vector in n-dimensional space that projects onto the real numbers $a_1, \ldots, a_n$ on the coordinate axes.

Quaternions

In 1843 the Irish mathematician-astronomer William Rowan Hamilton represented four-dimensional vectors algebraically and invented the quaternions, the first noncommutative algebra to be extensively studied. Multiplying quaternions with one coordinate zero led Hamilton to discover fundamental operations on vectors. Nevertheless, mathematical physicists found the notation used in vector analysis more flexible—in particular, it is readily extendable to infinite-dimensional spaces. The quaternions remained of interest algebraically and were incorporated in the 1960s into certain new particle physics models.

Projections. As readily available computing power grew exponentially in the last decades of the 20th century, computer animation and computer-aided design became ubiquitous. These applications are based on three-dimensional analytic geometry. Coordinates are used to determine the edges or parametric curves that form boundaries of the surfaces of virtual objects. Vector analysis is used to model lighting and determine realistic shadings of surfaces.

As early as 1850, Julius Plücker had united analytic and projective geometry by introducing homogeneous coordinates that represent points in the Euclidean plane and at infinity in a uniform way as triples. Projective transformations, which are invertible linear changes of homogeneous coordinates, are given by matrix multiplication. This lets computer graphics programs efficiently change the shape or the view of pictured objects and project them from three-dimensional virtual space to the two-dimensional viewing screen.

ALGEBRAIC GEOMETRY

Algebraic geometry is the study of the geometric properties of solutions to polynomial equations, including solutions in dimensions beyond three. (Solutions in two and three dimensions are first covered in plane and solid analytic geometry, respectively.)

Classification of curves by their genus

Algebraic geometry emerged from analytic geometry after 1850 when topology, complex analysis, and algebra were used to study algebraic curves. An algebraic curve C is the graph of an equation $f(x, y) = 0$, with points at infinity added, where $f(x, y)$ is a polynomial in two complex variables that cannot be factored. Curves are classified by a nonnegative integer—known as their genus g—that can be calculated from their polynomial.

The equation $f(x, y) = 0$ determines y as a function of x at all but a finite number of points of C. Since x takes values in the complex numbers, which are two dimensional over the real numbers, the curve C is two dimensional over the real numbers near most of its points. C looks like a hollow sphere with g hollow handles attached and finitely many points pinched together—a sphere has genus 0, a torus has genus 1, and so forth. The Riemann-Roch theorem uses integrals along paths on C to characterize g analytically.

A birational transformation matches up the points on two curves via maps given in both directions by rational functions of the coordinates. Birational transformations preserve intrinsic properties of curves, such as their genus, but provide leeway for geometers to simplify and classify curves by eliminating singularities (problematic points).

An algebraic curve generalizes to a variety, which is the solution set of r polynomial equations in n complex variables. In general, the difference $n - r$ is the dimension of the variety—*i.e.*, the number of independent complex parameters near most points. For example, curves have (complex) dimension one and surfaces have (complex) dimension two. The French mathematician Alexandre Grothendieck revolutionized algebraic geometry in the 1950s by generalizing varieties to schemes and extending the Riemann-Roch theorem.

Arithmetic geometry combines algebraic geometry and number theory to study integer solutions of polynomial equations. It lies at the heart of the British mathematician Andrew Wiles's 1995 proof of Fermat's last theorem.

(R.A.B./H.J.D'S.)

Non-Euclidean geometry

Non-Euclidean geometry is literally any geometry that is not the same as Euclidean geometry. Although the term is frequently used to refer only to hyperbolic geometry, common usage also includes those few geometries (hyperbolic and spherical) that differ from but are very close to Euclidean geometry (see Table 1).

The non-Euclidean geometries developed along two different historical threads. The first thread started with the search to understand the movement of stars and planets in the apparently hemispherical sky. For example, Euclid (*c.* 300 BC) wrote about spherical geometry in his astronomical work *Phaenomena*. In addition to looking to the heav-

Table 1: Comparison of Euclidean, Spherical, and Hyperbolic Geometries

Given a line and a point not on the line, there exists ___________ through the given point and parallel to the given line.	
(a) exactly one line	(Euclidean)
(b) no line	(spherical)
(c) infinitely many lines	(hyperbolic)
Euclid's fifth postulate is ______.	
(a) true	(Euclidean)
(b) true	(spherical)
(c) false	(hyperbolic)
The sum of the interior angles of a triangle ___ 180 degrees.	
(a) =	(Euclidean)
(b) >	(spherical)
(c) <	(hyperbolic)

ens, the ancients attempted to understand the shape of the Earth and to use this understanding to solve problems in navigation over long distances (and later for large-scale surveying). These activities are aspects of spherical geometry.

Euclid's parallel postulate

The second thread started with the fifth ("parallel") postulate in Euclid's *Elements*:

> If a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, will meet on that side on which the angles are less than the two right angles.

For 2,000 years following Euclid, mathematicians attempted either to prove the postulate as a theorem (based on the other postulates) or to modify it in various ways. These attempts culminated when the Russian Nikolay Lobachevsky (1827) and the Hungarian János Bolyai (1831) independently published a description of a geometry that, except for the parallel postulate, satisfied all of Euclid's postulates and common notions. It is this geometry that is called hyperbolic geometry.

SPHERICAL GEOMETRY

From early times, people noticed that the shortest distance between two points on Earth were great circle routes. For example, the Greek astronomer Ptolemy wrote in his *Geography* (*c.* AD 150) "it has been demonstrated by mathematics that the surface of the land and water is in its entirety a sphere . . . and that any plane which passes through the centre makes at its surface, that is, at the surface of the Earth and of the sky, great circles."

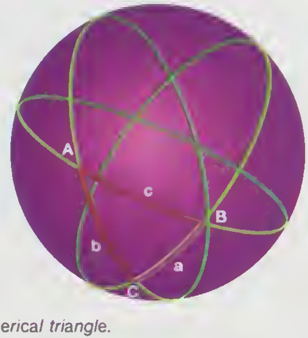


Figure 46: *Spherical triangle.*
The intersection of the three great circles, *AB*, *AC*, and *BC*, form the spherical triangle with sides a, *b*, and c on the sphere's surface.

Great circles are the "straight lines" of spherical geometry. This is a consequence of the properties of a sphere, in which the shortest distances on the surface are great circle routes. Such curves are said to be "intrinsically" straight. (Note, however, that intrinsically straight and shortest are not identical, as shown in Figure 42.) Three intersecting great circle arcs form a spherical triangle (see Figure 46); while a spherical triangle must be distorted to fit on another sphere with a different radius, the difference is only one of scale. In differential geometry (discussed above), spherical geometry is described as the geometry of a surface with constant positive curvature.

There are many ways of projecting a portion of a sphere, such as the surface of the Earth, onto a plane. These are known as maps or charts, and they must necessarily distort distances and either area or angles. Cartographers' need for various qualities in map projections gave an early impetus to the study of spherical geometry.

Elliptic geometry

Elliptic geometry is the term used to indicate an axiomatic formalization of spherical geometry in which each pair of antipodal points is treated as a single point. An intrinsic analytic view of spherical geometry was developed in the 19th century by the German mathematician Bernhard Riemann. It is usually called the Riemann sphere (see Figure 47) and is studied in university courses on complex analysis. Some texts call this study Riemannian geometry (as well as spherical geometry), but this term more correctly applies to a part of differential geometry that gives a way of intrinsically describing any surface.

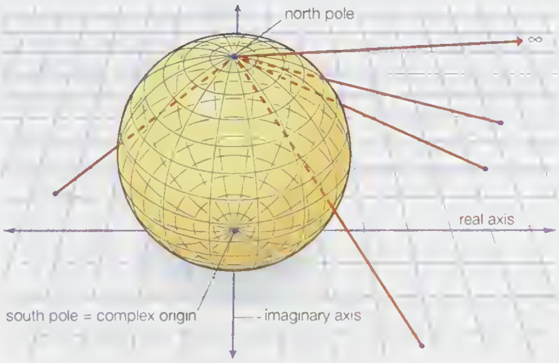


Figure 47: *Riemann sphere.*
With the south pole of the Riemann sphere placed on the origin of the complex plane (the intersection of the real and the imaginary axis), each line segment connecting points in the plane and the sphere's north pole intersects the sphere in a unique point. "Mapping" the points in the complex plane to the corresponding points on the sphere—and infinity to the sphere's north pole—transforms problems involving complex numbers into problems that reside in a portion of ordinary Euclidean space.

HYPERBOLIC GEOMETRY

The first description of hyperbolic geometry was given in the context of Euclid's postulates, and it was soon proved that all hyperbolic geometries differ only in scale (in the same sense that spheres differ only in size). In the mid-19th century it was shown that hyperbolic surfaces must have constant negative curvature. However, this still left open the question of whether any surface with hyperbolic geometry actually exists.

In 1868 the Italian mathematician Eugenio Beltrami described a surface, called the pseudosphere (shown in Figure 27), that has constant negative curvature. However, the pseudosphere is not a complete model for hyperbolic geometry because intrinsically straight lines on the pseudosphere may intersect themselves and cannot be continued past the bounding circle (neither of which is true in hyperbolic geometry). In 1901 the German mathematician David Hilbert proved that it is impossible to define a complete hyperbolic surface using real analytic functions (essentially, functions that can be expressed in terms of ordinary formulas). In those days, a surface always meant one defined by real analytic functions, and so the search was abandoned. However, in 1955 the Dutch mathematician Nicolaas Kuiper proved the existence of a complete hyperbolic surface; and in the 1970s the American mathematician William Thurston described the construction of a hyperbolic surface. Such a surface, as shown in Figure 48, can also be "crocheted."

In the 19th century, mathematicians developed three models of hyperbolic geometry that can now be interpreted as projections (or maps) of the hyperbolic surface. Although these models all suffer from some distortion—similar to the way that flat maps distort the spherical Earth—they are useful individually and in combination as aides to understand hyperbolic geometry. In 1869–71 Beltrami and the German mathematician Felix Klein devel-

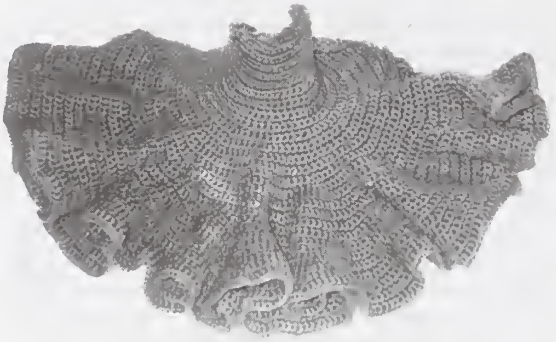


Figure 48: Hyperbolic plane, designed and crocheted by Daina Taimina.

By courtesy of Daina Taimina, Cornell University, Ithaca, New York

oped the first complete model of hyperbolic geometry (and first called the geometry "hyperbolic"). In the Klein-Beltrami model (see Figure 49, left), the hyperbolic surface is mapped to the interior of a circle, with geodesics in the hyperbolic surface corresponding to chords in the circle. Thus, the Klein-Beltrami model preserves "straightness" but at the cost of distorting angles. Around 1880 the French mathematician Henri Poincaré developed two more models. In the Poincaré disk model (Figure 49, right), the hyperbolic surface is mapped to the interior of a circular disk, with hyperbolic geodesics mapping to circular arcs (or diameters) in the disk that meet the bounding circle at right angles. In the Poincaré upper half-plane model (Figure 49, bottom), the hyperbolic surface is mapped onto the half-plane above the x-axis, with hyperbolic geodesics mapped to semicircles (or vertical rays) that meet the x-axis at right angles. Both Poincaré models distort distances while preserving angles as measured by tangent lines. (D.W.H./D.Ta.)

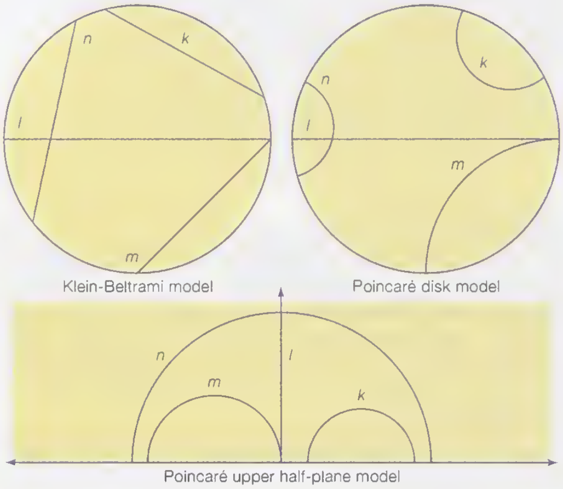


Figure 49: Three models of the hyperbolic plane.

BIBLIOGRAPHY

General history. The best overview in English of the history of geometry and its applications consists of the relevant chapters of MORRIS KLINE, *Mathematical Thought from Ancient to Modern Times* (1972, reissued in 3 vol., 1990), which can be supplemented, for further applications, by *Mathematics in Western Culture* (1953, reissued 1987). Three other useful books of large scope are PETR BECKMANN, *A History of π*, 4th ed. (1977, reissued 1993); JULIAN LOWELL COOLIDGE, *A History of Geometrical Methods* (1940, reissued 1963); and DAVID WELLS, *The Penguin Dictionary of Curious and Interesting Geometry* (1991). A fine recent survey at a college level of the various branches of geometry, with much historical material, is DAVID A. BRANNAN, MATTHEW F. ESPLEN, and JEREMY J. GRAY, *Geometry* (1999).

Ancient Greek geometry. The standard English editions of the Greek geometers are those prepared by THOMAS LITTLE HEATH beginning in the 1890s. They contain important historical and critical notes. Most exist in inexpensive reprints: *Apollonius of Perga: Treatise on Conic Sections* (1896, reissued 1961); *The Works of Archimedes* (1897, reissued 1953); *Aristarchus of Samos, The Ancient Copernicus* (1913, reprinted 1981); and *The Thirteen Books of Euclid's Elements*, 2nd ed., rev. with additions, 3 vol. (1926, reissued 1956). The historical material has been shortened and simplified, and its coverage extended, in *A History of Greek Mathematics*, 2 vol. (1921, reprinted 1993). Further information about technical-historical points—for example, the lunules of Hippocrates—may be found in WILBUR RICHARD KNORR, *The Ancient Tradition of Geometric Problems* (1986, reissued 1993). The epistemology of Greek geometry can be approached via the editor's introduction to and the text of Proclus, *A Commentary on the First Book of Euclid's Elements*, trans. and ed. by GLENN R. MORROW (1970, reprinted 1992).

Ancient non-Greek geometry. Other ancient geometrical traditions are covered in A.K. BAG, *Mathematics in Ancient and Medieval India* (1979); RICHARD J. GILLINGS, *Mathematics in the Time of the Pharaohs* (1972, reprinted 1982); JOSEPH NEEDHAM, *Mathematics and the Sciences of the Heavens and the Earth* (1959), vol. 3 of *Science and Civilization in China*; and B.L. VAN DER WAERDEN, *Science Awakening*, 4th ed., 2 vol. (1975).

Geometry in Islām. Aspects of the extensive development of geometry by Islāmic mathematicians can be studied in J.L. BERGGREN, *Episodes in the Mathematics of Medieval Islam* (1986). Otherwise, the best route to a survey is through the relevant chapters in vol. 2 of ROSHDI ROSHED (RUSHDI RASHID) (ed.), *Histoire des Sciences Arabes*, 3 vol. (1997), and the articles on Arab mathematicians and astronomers in CHARLES COULSTON GILLISPIE (ed.), *Dictionary of Scientific Biography*, 18 vol. (1970–90).

Renaissance geometry and applications. J.L. HEILBRON, *Geometry Civilized: History, Culture, and Technique* (1998, reissued 2000), considers examples of geometry from some modern cultures as well as from the ancient Mediterranean and gives examples of the development of Greek geometry in the Middle Ages and Renaissance. A more advanced book along similar lines, but with more restricted coverage, is ALISTAIR MACINTOSH WILSON, *The Infinite in the Finite* (1995). James Evans, *The History and Practice of Ancient Astronomy* (1998), is by far the best introduction to the theoretical and instrumental methods of the old astronomers. ALBERT VAN HELDEN, *Measuring the Universe* (1985), describes the methods of the Greeks and their development to the time of Halley. JOHN P. SNYDER, *Flattening the Earth: Two Thousand Years of Map Projections* (1993, reissued 1997), gives the neophyte cartographer a start. J.V. FIELD, *The Invention of Infinity: Mathematics and Art in the Renaissance* (1997), contains an elegant account, in both words and pictures, of the theory of projection of Brunelleschi, Alberti, and their followers. (J.L.He.)

Euclidean geometry. BENNO ARTMANN, *Euclid: The Creation of Mathematics* (1999), presents the contents of the *Elements* in modern terms accessible to a general reader and shows how many aspects of modern mathematics are prefigured by Euclid. H.S.M. COXETER, *Introduction to Geometry*, 2nd ed. (1969, reissued 1989), is a very readable scientific work starting from elementary Euclidean geometry and going on to more advanced topics. ROBIN HARTSHORNE, *Geometry: Euclid and Beyond* (2000), is an exhaustive modern presentation for mathematics students; it includes an extensive bibliography. EUCLID, *The Thirteen Books of Euclid's Elements*, trans. by THOMAS L. HEATH, 2nd ed., rev., 3 vol. (1926, reprinted 1956), is the standard English translation, with extensive commentary by Heath. DAVID HILBERT, *Foundations of Geometry*, 2nd ed. (1971, reissued 1992; trans. from German 10th ed., rev. and enlarged, 1968), gives a rigorous and logical account of the foundations of the subject. (B.Ar.)

Projective geometry. All of the following books contain some discussion of projective geometry by one of the greatest expositors of geometry, H.S.M. COXETER: with S.L. GREITZER, *Geometry Revisited* (1967), suitable for high school students; *Introduction to Geometry*, 2nd ed. (1969), suitable for advanced high school students and general undergraduates; and *Projective Geometry*, 2nd ed. (1974, reprinted with corrections, 1994); *The Real Projective Plane*, 3rd ed. (1993); and *Non-Euclidean Geometry*, 6th ed. (1998), all suitable for undergraduate mathematics students. (B.Ar.)

Differential geometry. JAMES CASEY, *Exploring Curvature* (1996), is a truly delightful book full of "experiments" to explore the curvature of curves and surfaces. JEFFREY R. WEEKS, *The Shape of Space: How to Visualize Surfaces and Three-Dimensional Manifolds* (1985), contains an elementary but deep discussion of different two- and three-dimensional spaces that may be models for the shape of our physical universe. ALFRED GRAY, *Modern Differential Geometry of Curves and Surfaces*

with Mathematica, 2nd ed. (1998), allows exploration of the subject through computer-generated figures. VLADIMIR ROVENSKI, *Geometry of Curves and Surfaces with MAPLE* (2000), is another textbook-software package for exploring differential geometry. DAVID W. HENDERSON, *Differential Geometry: A Geometric Introduction* (1998), emphasizes the underlying geometric intuitions, rather than analytic formalisms. JOHN MCCLEARY, *Geometry from a Differentiable Viewpoint* (1994), emphasizes the history of the subject from Euclid's fifth (parallel) postulate and the development of the hyperbolic plane through the genesis of differential geometry. RICHARD S. MILLMAN and GEORGE D. PARKER, *Elements of Differential Geometry* (1977), uses linear algebra extensively to treat the formalisms of extrinsic differential geometry. SAUL STAHL, *The Poincaré Half-Plane: A Gateway to Modern Geometry* (1993), is an analytic introduction to some of the ideas of intrinsic differential geometry starting from calculus. ETHAN D. BLOCH, *A First Course in Geometric Topology and Differential Geometry* (1997), explores the notion of curvature on polyhedra and contains the topological classification and differential geometry of surfaces.
(D.W.H.)

Analytic and algebraic geometry. CARL B. BOYER, *History of Analytic Geometry* (1956, reissued 1988), traces the early development of analytic geometry. JULIAN LOWELL COOLIDGE, *A History of Geometrical Methods* (1940, reissued 1963), provides proofs of important results in the history of analytic geometry. GORDON FULLER and DALTON TARWATER, *Analytic Geometry*, 7th ed. (1992, reissued 1994), is a classic introduction to the subject. ROBERT BIX, *Conics and Cubics: A Concrete Introduction to Algebraic Curves* (1998), provides a transition from analytic to algebraic geometry. PHILLIP A. GRIFFITHS, *Introduction to Algebraic Curves*, trans. from Chinese (1989), develops the topological and analytical properties of complex curves. Vol. 1 of IGOR R. SHAFAREVICH, *Basic Algebraic Geometry*, 2nd rev. and expanded ed., 2 vol. (1994; originally published in Russian, 1988), elegantly demonstrates the power of modern approaches to higher-dimensional algebraic geometry. ROBIN HARTSHORNE, *Algebraic Geometry* (1977, reprinted with corrections, 1997), is the best précis of Grothendieck's work in the foundations of algebraic geometry. SIMON SINGH, *Fermat's Enigma* (also published as *Fermat's Last Theorem*, 1997), gives a historical introduction to Fermat's last theorem and its proof.
(R.A.B./H.J.D'S.)

Non-Euclidean geometry. DAVID W. HENDERSON and DAINA TAIMINA, *Experiencing Geometry in Euclidean, Spherical, and Hyperbolic Spaces*, 2nd ed. (2001), compares non-Euclidean geometries and includes directions for constructing hyperbolic surfaces.
(D.W.H./D.Ta.)